z

Business Analytics: Data Visualization Project

Date: 12/12/2021

Title of the project: Analysis of US Accidents (2020)

Vertical industry: Mobility Industry

Business process: Safety, Administration and Prevention Measures

Submitted by: Shephali Jain

I certify that I have completed this assignment within the Academic Integrity guidelines presented in the UW General Catalog. Further, I certify that I do not have any knowledge of any other individual(s) violating these guidelines.

INDEX OF CONTENTS

INDEX OF TABLES

1 Analysis of US Accidents (2020) – Final Project

1.1 Executive Summary

Reducing traffic accidents is an essential public safety challenge all over the world; therefore, accident analysis has been a subject of much research in recent decades. The objective of the project is to perform exploratory and diagnostic analysis of the changes in year 2020 to analyze the conterminous [1]US accident data from 49 states (based on a data set obtained from Kaggle, US Accidents), and to inform the US government agencies possible causes of traffic accidents since Covid-19 hit, specifically for year 2020 and what could be done to reduce them.

During this time, the pandemic due to the coronavirus (SARS –CoV-2) was a threat to the health and welfare of all people residing in US. Government issued several executive orders that would help to slow the spread of the virus, and its associated health impacts now commonly known as COVID-19. The analysis include number of accidents by year, number of accidents by state, best time to travel by month, day and hour, accident-prone area in each state, factors responsible of the accidents like weather, wind flow, temperature, location, etc.

Business Challenge Questions- As Covid-19 affected the public transports over the years 2019-2020, majority of population is bound to arrange the travel medium on their own resulting into increase in purchase of vehicles, causing more possibilities for accidents to occur. In order to minimize the economic cost associated with crashes and to reduce life loss, a deep analysis of mentioned factors and parameters are important as Government consideration.

Government agencies and the general public can leverage these insights from the analysis and take a preventive measure which can reduce US accidents based on the following questions answered:

1. What are the top accident-prone areas in the US (State and city)?
2. What is happening in the highest accident-prone states and cities, what are the causes?
3. What day and time are safe to travel?
4. What are the factors responsible for accidents?
5. What is the severity of these accidents?
6. What solution can be implemented to reduce accidents?
7. How can these accidents be minimized to reduce economic cost of motor vehicle crashes?

[1] The contiguous United States or officially the conterminous United States consists of the 48 adjoining U.S. states and the District of Columbia on the continent of North America. The terms exclude the non-contiguous states of Alaska and Hawaii and all other offshore insular areas, such as American Samoa, Guam, the Northern Mariana Islands, Puerto Rico, and the U.S. Virgin Islands.

1.2 Background

1.2.1 Introduction

Car accidents are unexpected events that occur to motor vehicles causing damage of the vehicles, structures, fatalities and even death of the people in the vehicles. According to research done by the United Nations, the rate of growth of accidents in the world continue to increase with over two million deaths and thirty million injuries reported annually.

In the United States and throughout much of the world, car accidents are a leading cause of serious injury and death. In fact, in the U.S. alone, at least 38,800 people were killed in motor vehicle collisions in 2019 generating economic cost of crashes approx. $242 billion. Despite the events of 2020 and the response to the Coronavirus pandemic, initial reports have indicated that motor vehicle fatalities actually increased in the United States in 2020 over 2019. Clearly, car and motor vehicle safety is still a top issue for all people.

National Highway Traffic Safety Administration (**NHTSA**) is an agency which is part of the United States Department of Transportation tasked with upholding regulatory safety standards in automobile manufacturing and highway safety.

The numbers of vehicles continue increasing every year and governments are responding by constructing modern roads that can facilitate smooth transportation of goods and people so as to realize economic growth and reduce accidents which are increasing dramatically. In fact, car accident are ranked second to the major killer disease AIDS in terms of causing deaths and lose of property and resources. As a result, government policies to increase economic empowerment among the people are destructed by car accidents.

After World War II, the automobile engine picked up the preeminent position as the primary means of transport. From that point forward, no one has challenged the dominance of engine vehicles. Instead, there have been various efforts to improve them, for example: to make the assembly line faster, to make them progressively adapted to the geographical terrain found in individual nations.

At present, automobile transport has become a piece of daily life. Improvement of automobiles is inescapable given the shockingly on-going high rates of terrible accidents and deaths. Unfortunately, vehicle crashes have always been a part of the vehicle driving experience.

Over 1.2 million individuals die every year on the world's streets, and somewhere in the range of 20 and 50 million endure non-fatal injuries. To show the significance of traffic accidents globally, the World Health Organization (WHO), in its worldwide status report on road safety 2009, estimates that in high income nations like the USA there are 65 % of reported vehicle deaths from the Vehicle Occupants as compared to middle-income countries of the western pacific locale where 70% of the deaths are among vulnerable street users (WHO,2009). The same report also predicts that road traffic injuries will rise to become the 5th leading cause of death by 2030 (WHO,2009).

Although the global loss and suffering resulting from road accidents are indeed small compared with that caused by poverty and sickness, the problem is more severe than the present figures alone indicate. It is necessary to consider the monetary loss to nation-states due to fatal automobile accidents. A large number of the fatalities happen indiscriminately to vehicle users. In 2010, the economic loss of the USA alone was about 836 billion. (U.S. Department of Transportation,2015).These include educated individuals; the statesmen, specialists, instructors, and businesspeople whose loss to the nation is severe.

1.2.2 Desk Research

There is a great deal of research out there that addresses accidents occurring the world over (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv Ramnath,2019). Regardless of all these progressing research numbers of accidents happening are not decreasing, which is a primary worry to everybody. However, the vast majority of them are on accident analysis.

In one of the research papers "A Countrywide Traffic Accident Dataset" (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv Ramnath,2019), they have tried to address this issue by collecting the data from API resources available from various sources and having records of 2.25 million instances of traffic accidents that took place within the contiguous the United States, and over the last three years. Each accident record consists of a variety of intrinsic and contextual factors such as location, time, natural language description, weather, period-of-day, and points-of-interest (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv Ramnath,2019). Chang et al. (Chang, 2005) utilized information such as road geometry, annual average daily traffic, and weather data to predict the occurrences of accidents for a highway road by designing a neural network model.

Over time numerous studies have used large scale datasets; however, the datasets have been either private or not easily accessible (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv amnath,2019). Eisenberg (Eisenberg, 2004) carried analysis to identify the impact of road accidents with a large dataset of about 456000 crashes in 48 US states from 1975 to 2000. Recent studies by Najjar et al. (Najjar, Kaneko, & Miyanaga, 2017) have used large scale datasets to analyze real-time traffic accident prediction. Despite all these studies, results were not available for further research. The main thing about the dataset is, it is available publicly; however, it is limited in terms of one city or state, attributes are not enough for analysis.

Most of the research is not readily available to Government agencies and the public. If we consider all this research, we can find that there is a big gap between the result found from this research and the implementation of this outcome.

To address this challenge, we propose a new platform which can showcase all the finding by each state like day and time safe to travel, accident prone area and zip code in each state, severity, weather conditions, also if someone wants to go from Los Angeles to San Francisco in which area accidents mostly occur. For State Government officials, this platform will help to decide and provide solution-based on accident issues face by each state.

1.3 Learning Objectives

The following are the main learning objectives for this tutorial:

1. **Data Exploratory Analysis**- The research on the existing variables for the dataset will be performed based on the visual analysis to find out most relevant and significant attributes for answering the business problem via the tableau dashboard.

➔ An in-depth study for the datapoints indicating impactful variable significance toward the roots of the business problem are to be analyzed and would be taken as an input for further high-end analysis.

2. **Descriptive Analysis**- The analysis will be carried out to understand the 5 elementary Ws of the problem i.e. What, Where, When ,Why and Who via the metrics of the dependent factors represented in tableau story board.

➔ Based on the insights drawn from different combinations of attributes, the purpose of analysis will get directions for proposing the recommendations and suggestions to diagnose the issues stated.

3. **Diagnostic Analysis**: This analysis will unfold the hidden root cause of the problem at granular levels of detail of data processing. Demonstration of different charts used in the dashboards and storyboards will help deduce the recommendations for Government authorities and official departments for travel and safety related measures.

   Note- Using different charts for different type of data is a niche skill which helps in delivering stories in a more succinct and impactful way. Remembering a story is easier than remembering few metrics of the dashboard.

Above stated learning objectives will be covered throughout the documentation.

1.4 Methods & Description of Data

1.4.1 Methodology

The methods used to address the business problem are categorized in following parts:

A. Assessment of current accident situation in the United States with main focus on stats obtained for year 2020. In this part, we will be performing analysis to gain understanding of main questions like distribution of accidents according to the severity, time of the day, day of the month and month of the year 2020 safe to travel, and the rest overview.

B. Assessment of accident issues concerning topmost accident-prone states of the US and the topmost accident-prone cities of those states by zip code.

C. Additionally, the insights drawn from above can be taken forward to be shared for Government to take specific measures to avoid such terrible accidents and hazards resulting in human suffering. Possible suggestion on how these accidents can be reduced, and what are solutions can be implemented based on the crucial factors identified causing huge number of accidents.

1.4.2 Data Sources and Data Collection

A. Content Description

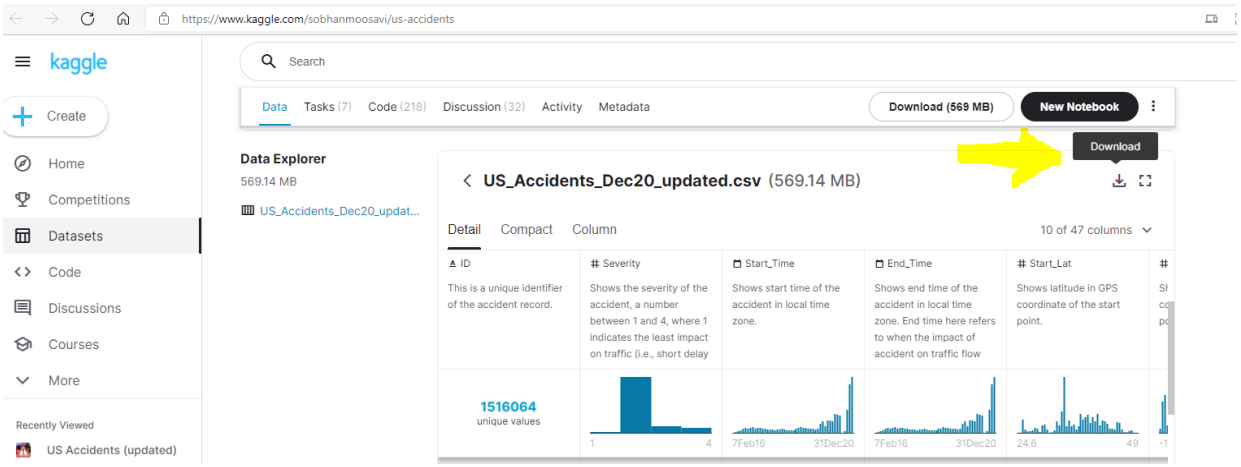
The data is continuously being collected from February 2016, using several data providers, including multiple APIs that provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about **1.5 million** accident records in this dataset along with 47 columns.

Dataset Source: Kaggle.com

Download Link: US Accidents (updated) | Kaggle

B. Steps to retrieve the dataset

The dataset can be downloaded in zip file and extracted by unzipping.



The format of the file is **.csv** i.e., a text file which can be easily connected to tableau.

Note: The dataset has been analyzed to be consistent with the specified criteria i.e., 40,000 rows and 7-10 columns.

C. Applications of Dataset

US-Accidents can be used for numerous applications such as real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence. The most recent release of the dataset can also be useful to study the impact of COVID-19 on traffic behavior and accidents.

D. Data Removal (Updated Dataset)

A portion of data was removed from this dataset due to a request from one of the main traffic data providers. Due to the removal, the dataset contains data only for Nov-Dec 2019 and year 2020 as shown in the below figure.

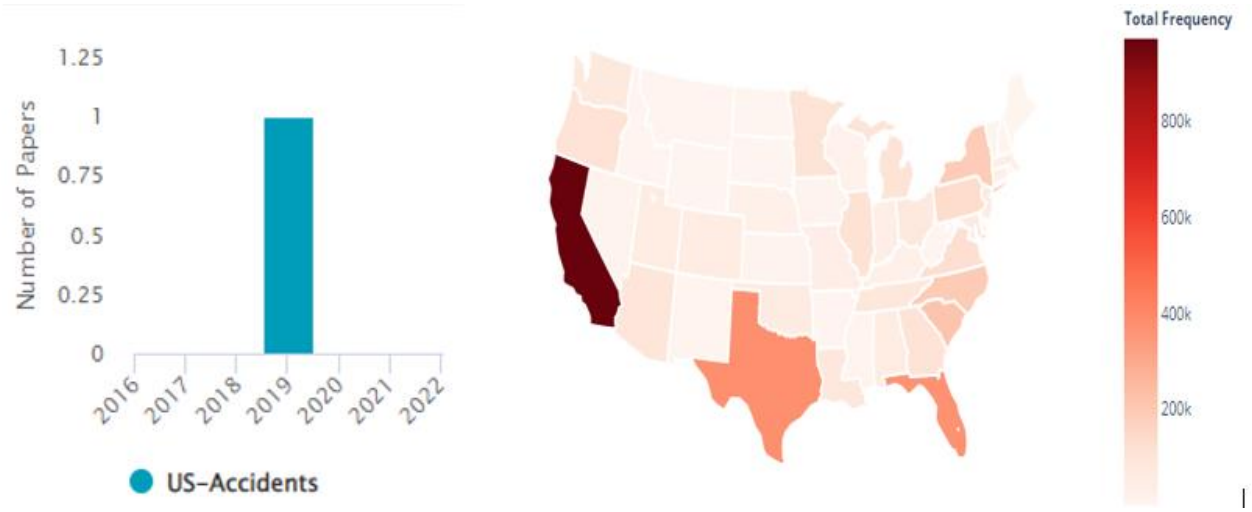


Figure-1 Data Availability by Year for US Accidents dataset and Frequency distributions of US-Accidents (Dec 2020)

The main problem statement aligns for the year 2020 accident count, so the limited counts of accident for year 2019 are not taken into consideration.

1.4.3 Data Dictionary

There are 47 attributes along with the field description as shown in below Table.

#	Attribute	Description
1	ID	This is a unique identifier of the accident record.
2	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).
3	Start_Time	Shows start time of the accident in local time zone.
4	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.
5	Start_Lat	Shows latitude in GPS coordinate of the start point.
6	Start_Lng	Shows longitude in GPS coordinate of the start point.
7	End_Lat	Shows latitude in GPS coordinate of the end point.
8	End_Lng	Shows longitude in GPS coordinate of the end point.
9	Distance(mi)	The length of the road extent affected by the accident.
10	Description	Shows natural language description of the accident.
11	Number	Shows the street number in address field.
12	Street	Shows the street name in address field.
13	Side	Shows the relative side of the street (Right/Left) in address field.
14	City	Shows the city in address field.
15	County	Shows the county in address field.
16	State	Shows the state in address field.
17	Zipcode	Shows the zip code in address field.
18	Country	Shows the country in address field.
19	Timezone	Shows time zone based on the location of the accident (eastern, central, etc.).
20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.
21	Weather_Timestamp	Shows the timestamp of weather observation record (in local time).
22	Temperature(F)	Shows the temperature (in Fahrenheit).
23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
24	Humidity(%)	Shows the humidity (in percentage).
25	Pressure(in)	Shows the air pressure (in inches).
26	Visibility(mi)	Shows visibility (in miles).
27	Wind_Direction	Shows wind direction.

28	Wind_Speed(mph)	Shows wind speed (in miles per hour).
29	Precipitation(in)	Shows precipitation amount in inches if there is any.
30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.
32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.
33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.
34	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.
35	Junction	A POI annotation which indicates presence of junction in a nearby location.
36	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.
37	Railway	A POI annotation which indicates presence of railway in a nearby location.
38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.
39	Station	A POI annotation which indicates presence of station in a nearby location.
40	Stop	A POI annotation which indicates presence of stop in a nearby location.
41	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.
42	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby loction.
43	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.
44	Sunrise_Sunset	Shows the period of day (i.e., day or night) based on sunrise/sunset.
45	Civil_Twilight	Shows the period of day (i.e., day or night) based on civil twilight.
46	Nautical_Twilight	Shows the period of day (i.e., day or night) based on nautical twilight.
47	Astronomical_Twilight	Shows the period of day (i.e., day or night) based on astronomical twilight.

Table 1-1. Data Dictionary Table for US-Accidents Dataset

1.4.4 Data Visualization Tool- Tableau

A. Why Tableau?

Tableau is a power BI Tool used in business organizations as it has many useful and compact functionalities which allow to present the data by slicing and dicing in proper manner asper the user report requirements. Below table represents the benefits of using Tableau over traditional reporting tools:

Traditional Method	Tableau
Prior programming skills	No programming skills required
Focused on only one type of database	Combines different types of database spreadsheets, databases, cloud data, and even big data such as Hadoop
Decision-makers have to ask the IT people to retrieve any information from the database	Decision-makers can directly use the dashboard to retrieve any information from the database
Mostly depends on Query languages	The query is done behind the scene
Combining different types of the database is difficult	Different types of databases can be combined easily
Not every business intelligence tool offers an interactive dashboard	The interactive dashboard is easy to build, and it makes data visualization quick and efficient
Mostly designed for large businesses	Perfect BI solution for small, medium, and large businesses, and even for non-profits
Comparatively expensive	Comparatively affordable
Time-consuming	Time-saving

Table 1-2. Benefits of Using Tableau for Data Visualization

1.5 Data Cleaning (Tableau Prep Builder)

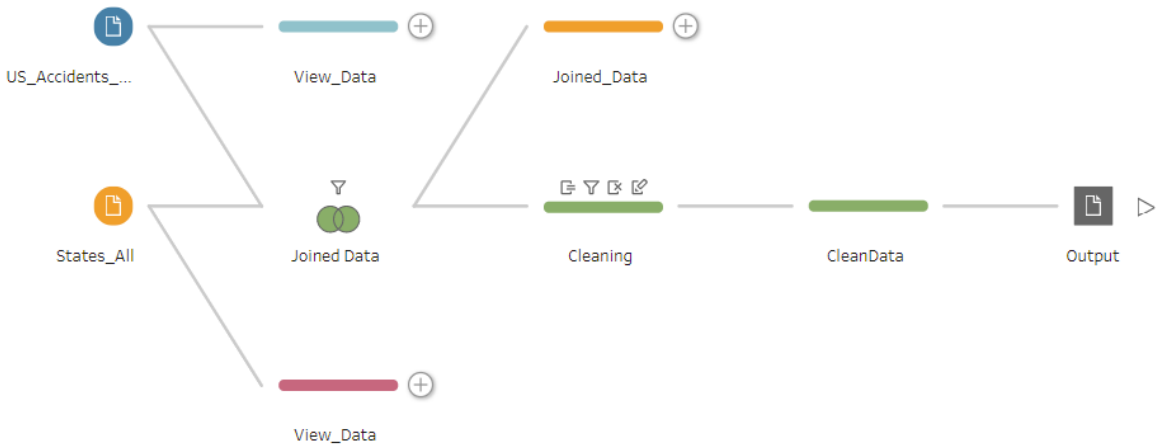
1.5.1 Identification of Anomalies

While the data set has 1.5 million records, it is not ready to use for analysis. There are many anomalies in the dataset listed below. To handle all these anomalies in data, data cleaning is the most important and mandatory step. The below steps to eliminate the noisy data are performed and presented below with screenshots for the dataflow and to produce the clean file as output to be consumed in Tableau.

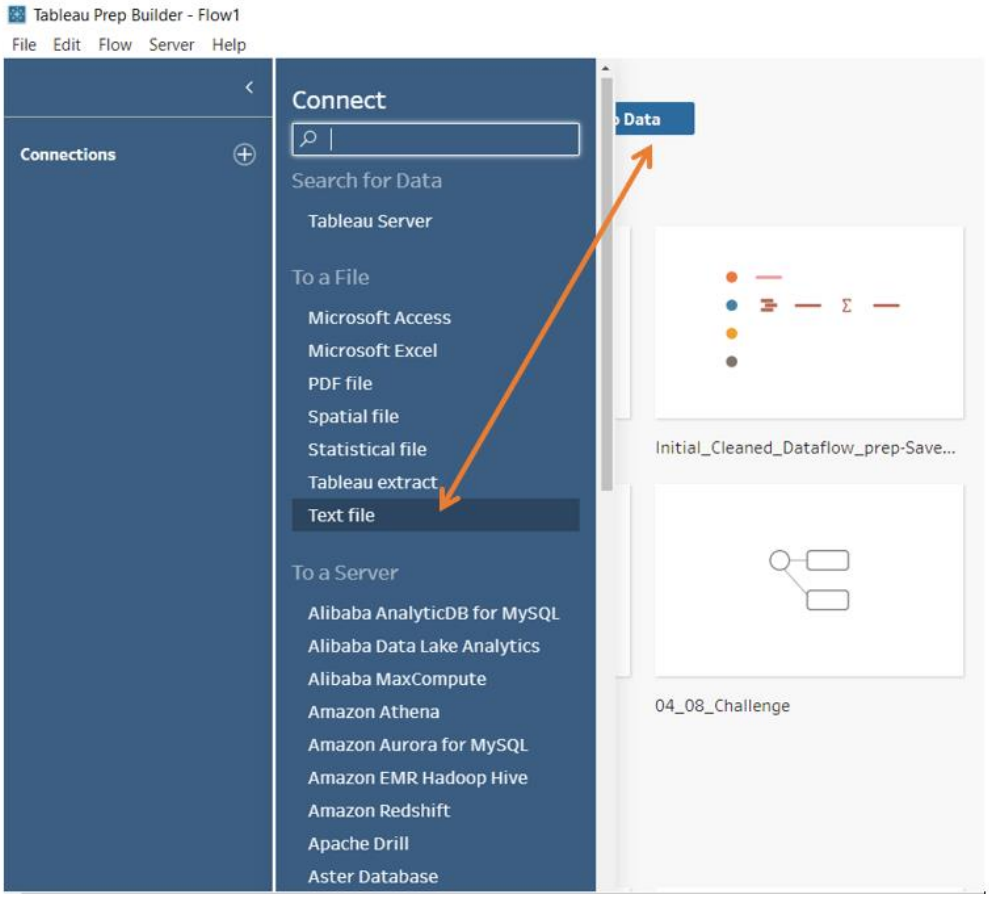
- ➔ **Null records**
  As the data had huge amount of null values, this was carefully assessed as per individual category and then removal of null values was performed as part of data cleaning.
- ➔ **Mismatched column**
  The auto-detection feature of tableau when the data is inserted, helps to identify correct datatypes but sometimes this acts as the limitation for few attributes. Correct data types were assigned manually.
- ➔ **State name missing**
  In the given dataset, only State codes are there without the elaborated state names. We have added the names of the states by performing join operation in tableau prep.
- ➔ **Feature Engineering for extracting "Day"**
  As "day of the accident" is needed to analyze the frequency of accidents and patterns followed in a week to know which day there is a raise in occurrences of accidents, "Start_Time" attribute has been used.
- ➔ **Unclean Zip Codes**
  Cleaning step with writing the below formula was done to obtain the cleaned Zip-codes (removal of area codes was done)

1.5.2 Cleaning Steps

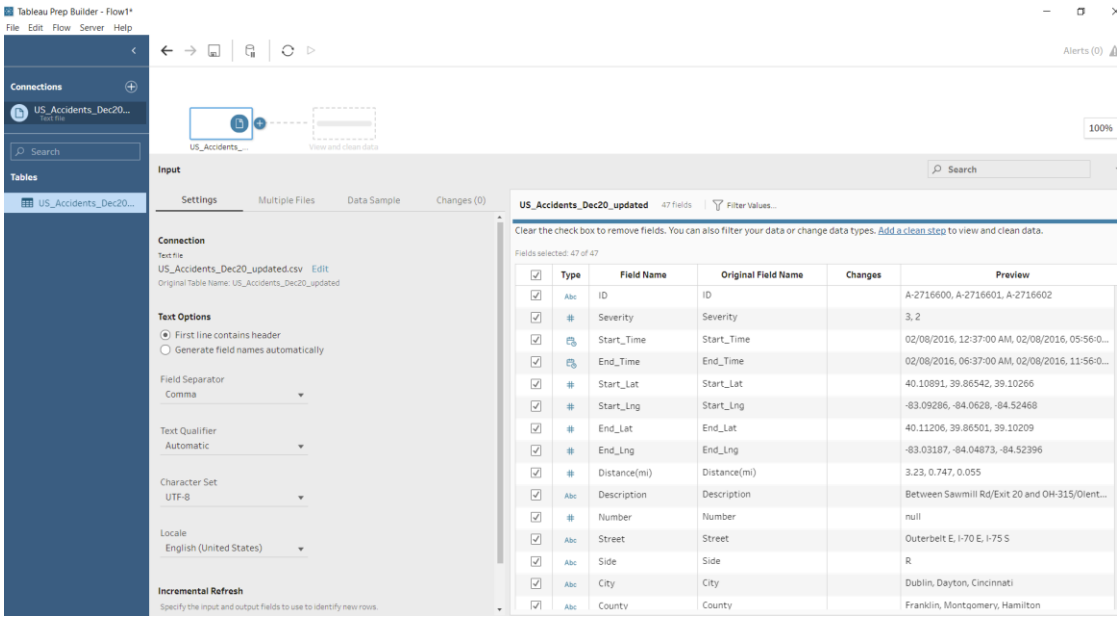
Steps performed for generating the cleaned dataset csv file via the below shown dataflow:



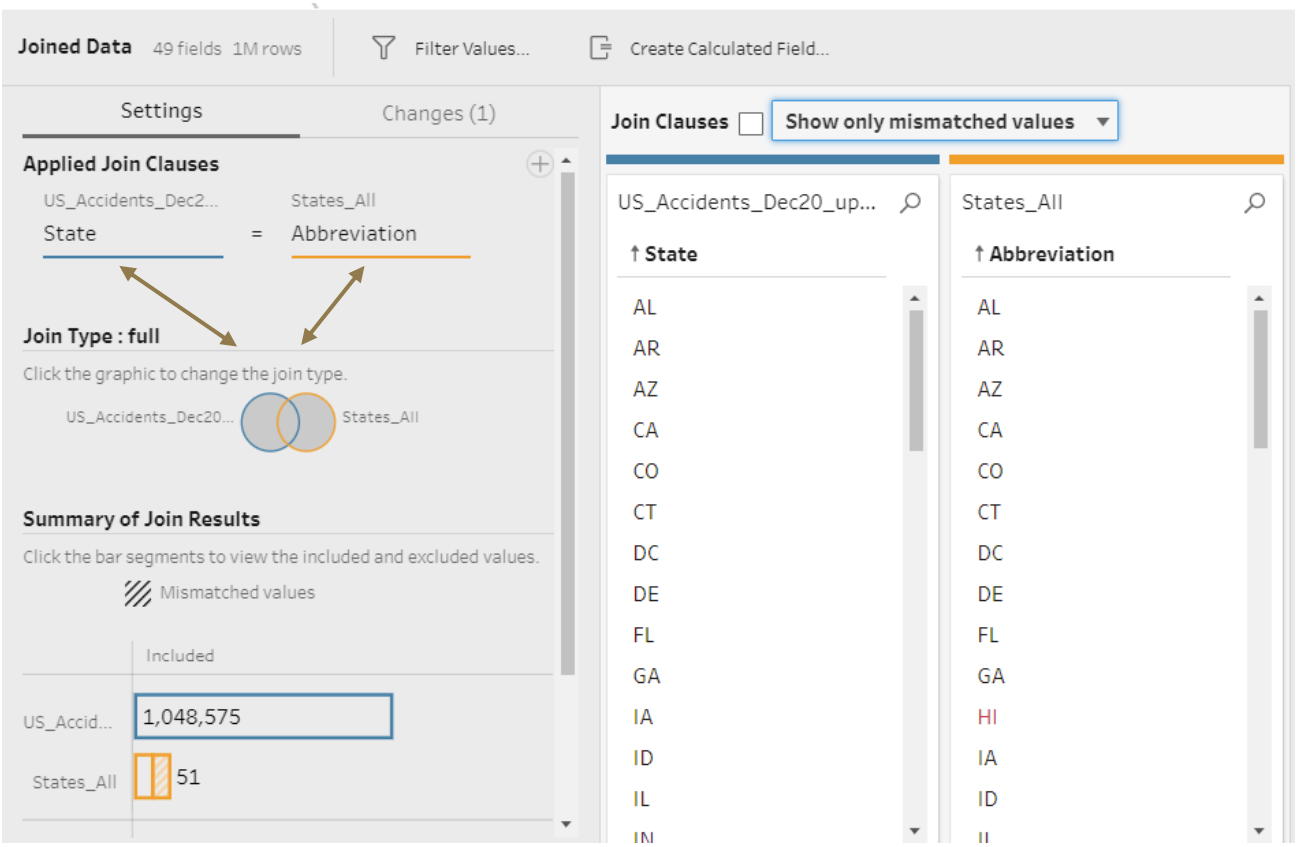
1. Added the raw US_Accidents_updated.csv file by connecting to data source option:



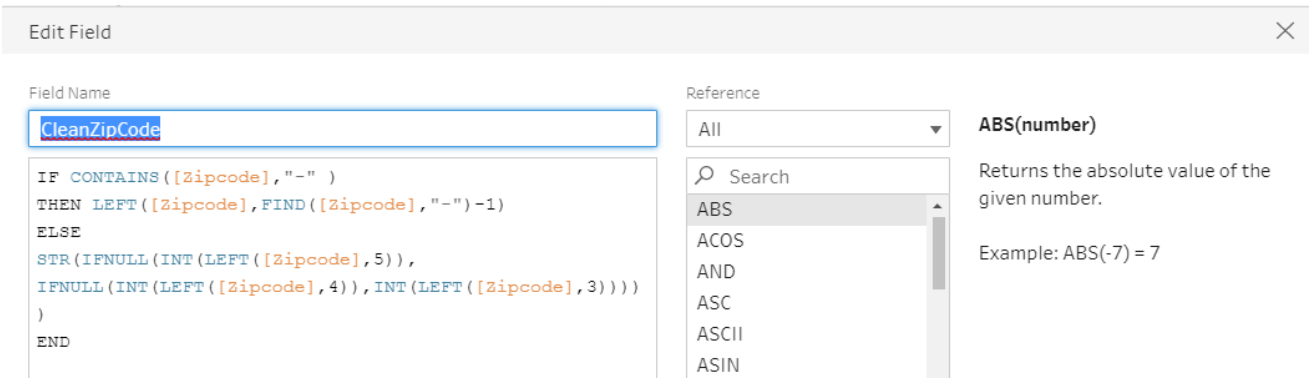
2. Next, we checked all the attributes and the associated data types:



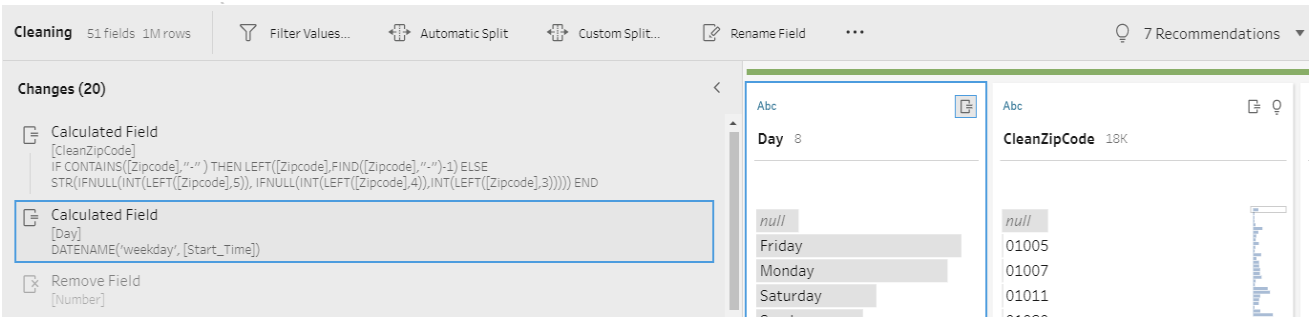
3. Since the dataset did not have state names elaborated (full forms with the abbreviations), we added the descriptive field of the states with another csv file joining both in tableau prep.



4. Cleaned the Zipcode by performing filter operation:



5. Extracted “Day” field from “Start_Time” using DATENAME function as shown in the screenshot below:



6. Removal of unnecessary columns
   Considering the percentage of null records, eliminated the following columns: **Number**

   As shown in the above screenshot, overall, 20 changes were made at the cleaning pill of tableau prep to obtain the cleaned data set as "Output_extract.csv" as shown below:

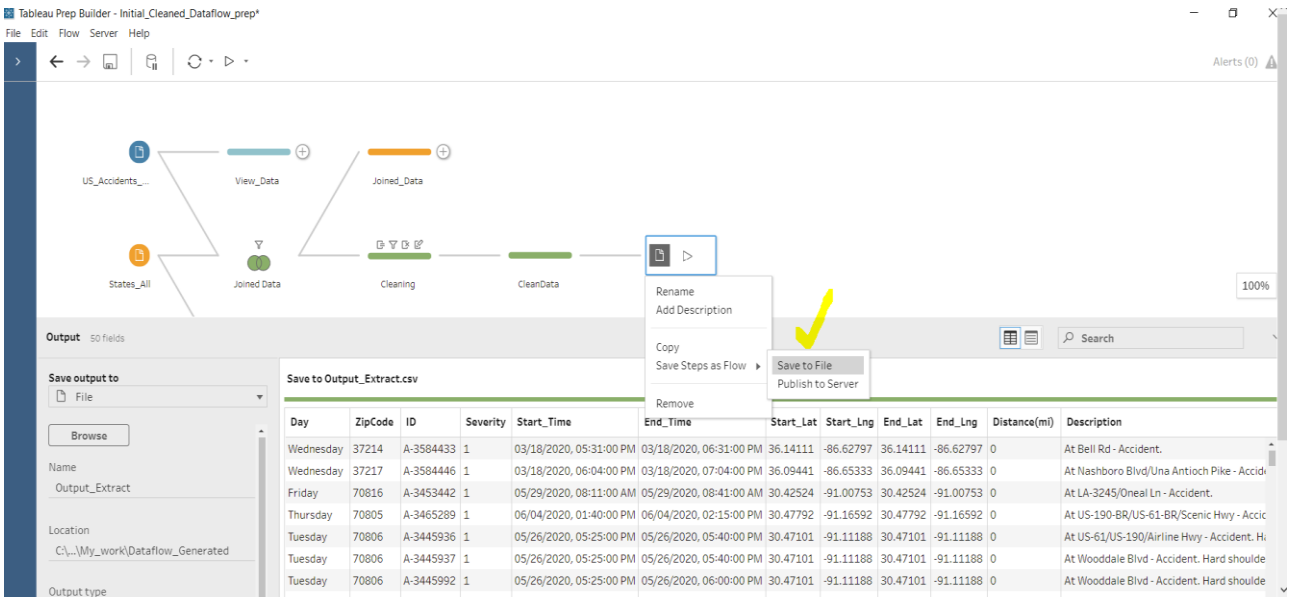


Figure- Final Dataflow Extract File.

Characteristics of traffic accidents were obtained from analyzing the available data. The data was collected, clean, manipulated, tabulated, and then analyzed.

1.5.3 Transformed dataset

The final dataset consists of 7,21,000 rows and 50 columns (including the newly created columns needed for analysis).

1.6 Deliverable 1 – Dashboard

1.6.1 Business Scenario

"Significant Factors Affecting Accident Rates by severity in the US (2020 Stats)"
To find out why there is so much increase in accidents in year 2020, it is required to start asking the right questions about which are the factors most affecting in raising accidents. In order to direct these questions, exploration to the dataset is a key step which builds the basic foundation of the analysis by data understanding and choosing the most significant features to look into.

Government and Transport department can use the list of elements such as multiple questions that are relevant to understand the factors behind the increase in accidents as an input in order to find a solution to those pain-points for infrastructure loss, economic cost, and human life loss.

The data consists of 50 attributes which are related to the accident data collected in all over the country to determine the causality relation among them and the responsible factors resulting into accidents occurrences. Summarizing these responsible factors as per the severity level will help in taking measures to reduce the delay caused by the accidents over the roads and consequently helping in minimizing the problem of congestions over the roads and highways. Security is one of the major concerns for the Government as a single accident occurrence might result into many due to traffic congestions.

Link: https://public.tableau.com/views/USAccidentsAnalysis2020/Storyboard?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

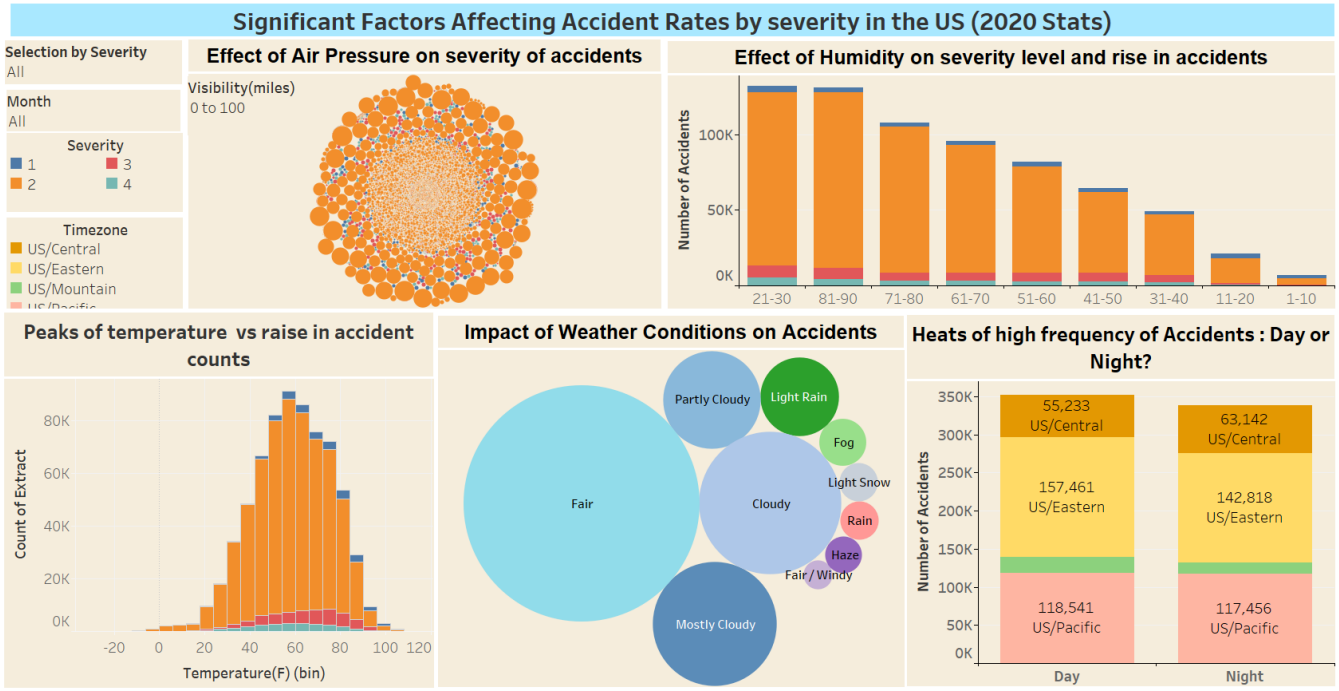


Figure- Dashboard

1.6.2 Applicable Questions

The purpose of this business scenario is highly dependent on the relevant attributes selected to calculate the significance among all given in the dataset which is part of the descriptive analysis.
We have used the Data processing measures by running regression models to find the top relevant factors included in the recorded steps in Procedure section of the documentation.

Noteworthy Purposes of performed Descriptive-Diagnostic Analysis:
Q1. Accidents caused by different weather conditions in different months of the years in the US.
Q2. Accidents caused by the severity of levels and visibility in the US.
Q3. Accidents caused by the percentage of humidity and air pressure rise.
Q4. Accidents distribution over the day and night conditions.
Q5. Accidents distribution by the temperature recorded in Fahrenheit.

The above listed question are important for Government and Transportation Administrative Department to uncover the problem areas and to take corrective measures and precautions depending on the insights gained through the analysis.

By slicing and dicing the data through the filters of different combinations of attributes, new discoveries will be explored which are hidden in the raw data of US Accidents collectively.
As the accident rate has been increased from 2019 to 2020 by huge percentage, the authorities are dependent on these questions to implement the corrective measures specific to applicable scenarios relevant to different concern areas.

1.6.3 Procedure (Analysis & Visualizations)

Prerequisite step: Data processing to gather the most significant variables among the 51 attributes of the dataset and we will be using these into our analysis for finding insights relating to accident count.

Variables	Significance
Pressure (in)	12.20%
Weather_Conditions	10.50%
Humidity (%)	10.34%
Temperature (F)	3.55%
Day	9.18%
Traffic_signal	5.67%
Hours	8.03%
Wind_Speed	3.64%
Stop	4.23%
Junction	4.09%
Month	4.79%
Crossing	3.39%
Visibility(mi)	2.31%

Table 1-4 List of Significant attributes

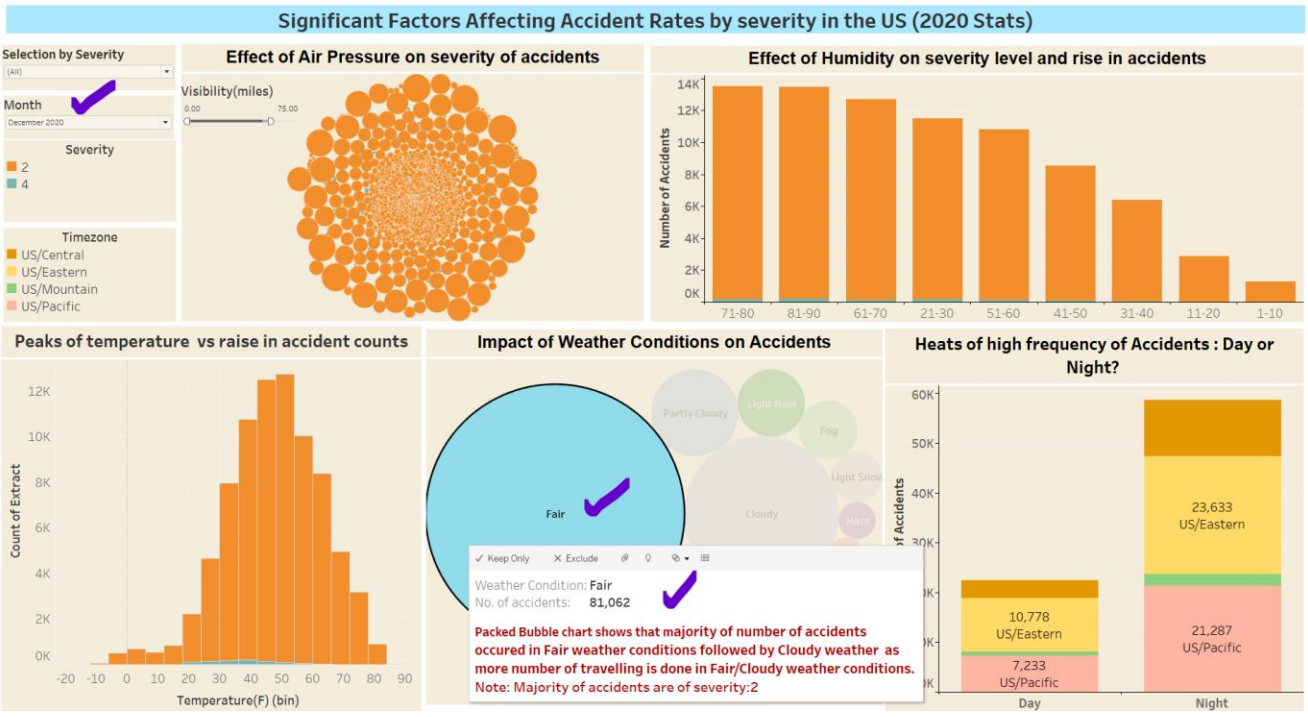
Tableau Tutorial

A dashboard is a type of graphical user interface which often provides at-a-glance views of key performance indicators relevant to a particular objective or business process.

Note: The filter list dropdown are applicable on all the visualizations presented in the dashboard and additionally, the viz elements can also be used to uncover the hidden information depending on change of filters.

Q1. Accidents caused by different weather conditions in different months of the years in the US.

Let's say if there is a need to check for December month of year 2020, how many accidents happened in Fair weather conditions. Using the Left topmost corner filter settings, we can select the Month and then clicking on "Fair" packed bubble will display the results containing number of accidents.



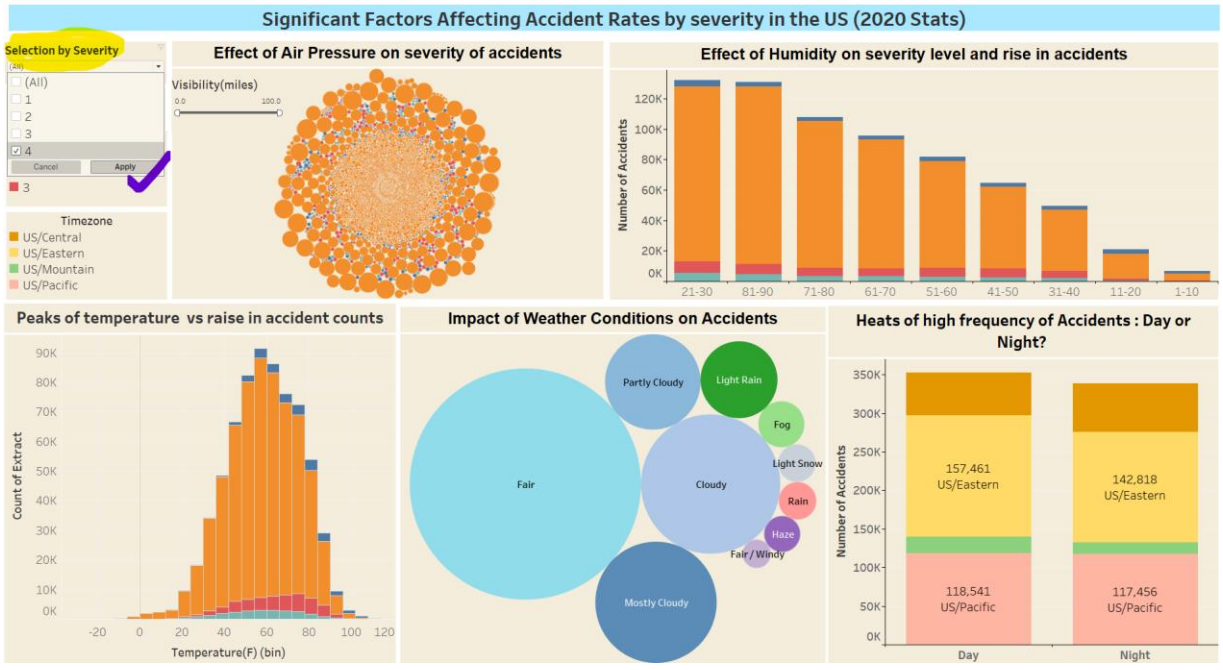
Insights Drawn:

- ➔ Notice that the accidents belong to severity 2 (causing short delay in traffic) and 4 (causing very high delay in traffic) in Fair weather category.
- ➔ Majority of the accidents belong to severity level 2 in other charts such as humidity % and Temperature peaks for the given filter criteria of December 2020.
- ➔ For Dec 2020, more accidents happened at nighttime and majority of the accidents fall in the Pacific and Eastern US Time zones.
- ➔ This is useful for determining the scenarios where administration officers for travel department needs to find out how the quarter wise implementation of diversified rules should be defined. For example, in different US Zones, depending on the weather conditions (Here we have taken top 10 weather conditions which are causing higher accidents) throughout the year, when there should be more awareness guidelines and traffic congestion management.

Q2. Accidents caused by the severity of levels and low visibility conditions in the US (2020).

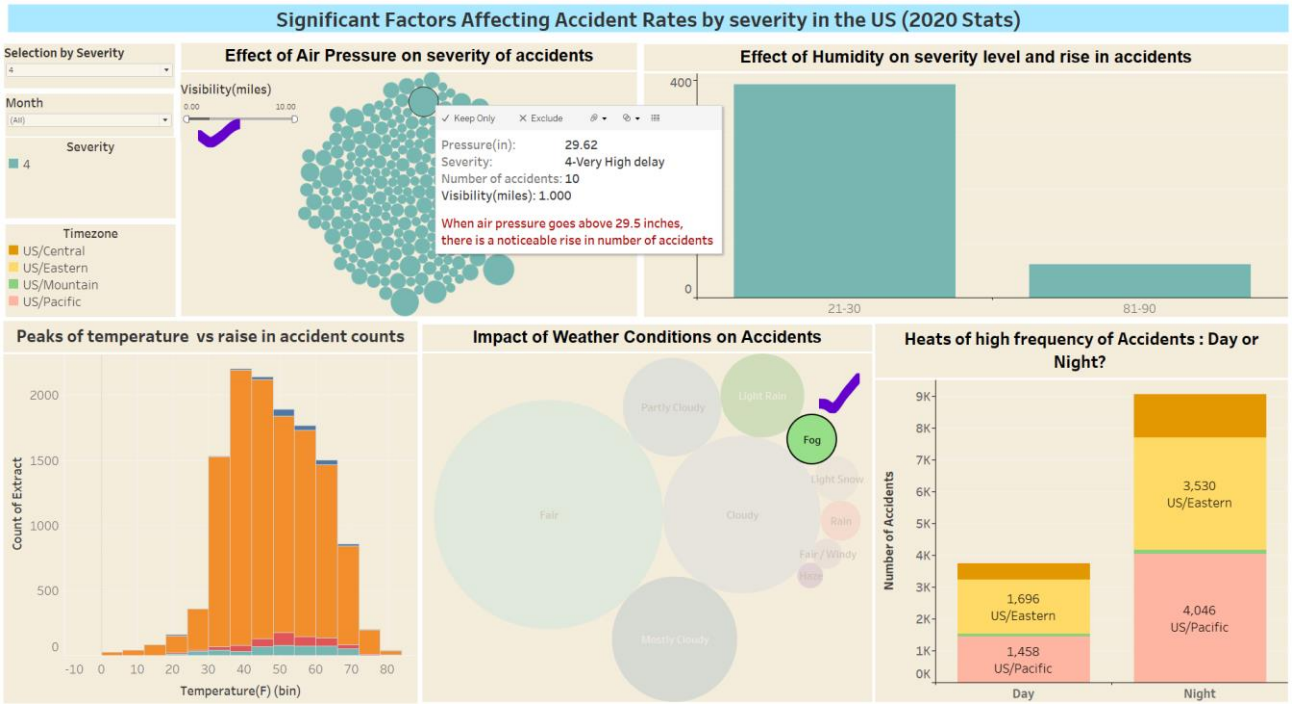
To find the number of accidents filtered based on severity = 4 (causing very high delay in traffic) and having low visibility (range of visible distance less than 10 miles on the road ahead) specially when weather conditions are defined with "fog".

A. Click on the filter to select **Severity = 4** from drill down and Click on **Apply**:



The 'Apply' filter will first show only Severity 4 Applicable accidents in all the visualizations.

B. Select the Visibility range from 0 – 10 miles using the slider and click on "Fog" packed bubble:



Insights Drawn:

➔ We can see the lowest visibility range (less than 10 miles specifically **1 mile** in the above case) distribution of accidents with severity 4 are displayed. Also, it is noteworthy to see that in Fog conditions, effect of Air pressure is above 29.62 inches.

➔ We can separately analyze the temperature peaks for Severity 4 accidents as they belong to all different bins of Temperature.

➔ For the applicable filters, the bar chart representing frequency of accidents during the night are more and Pacific and Eastern regions are again exposed to high number of accidents in US.

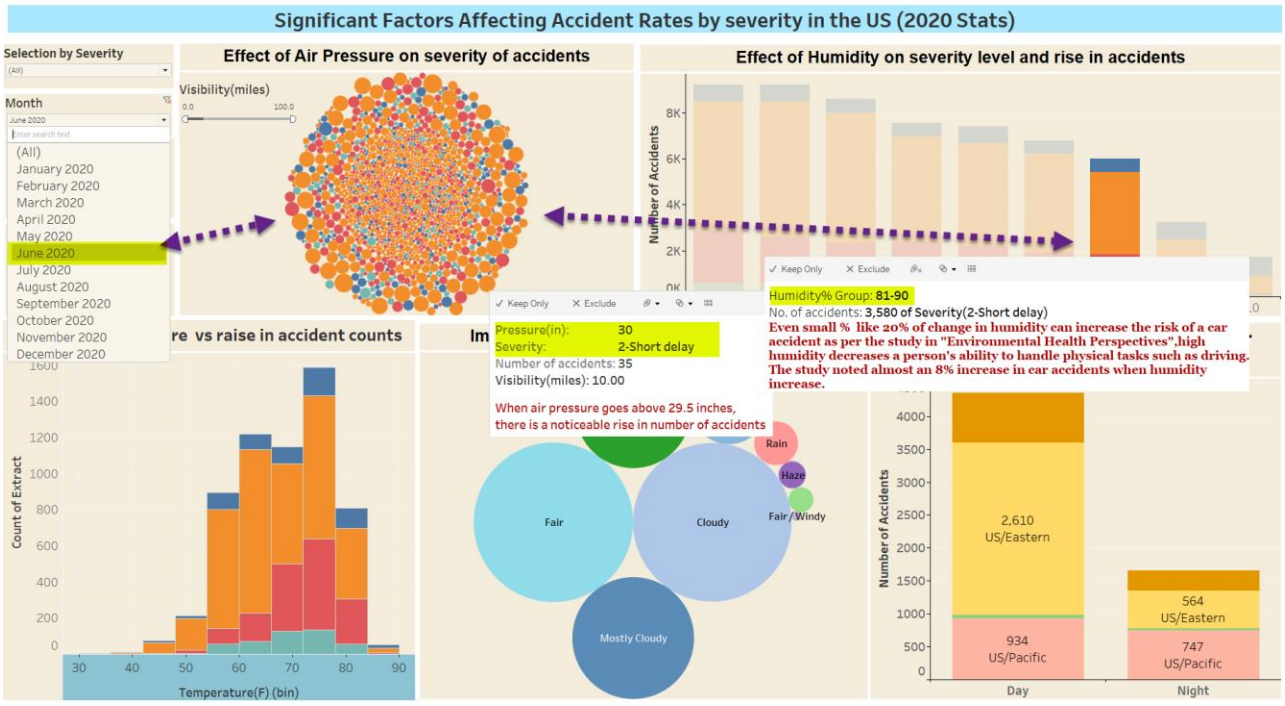
➔ Almost 80% of accidents occurred in the presence of Fog are under the severity 2 and it is an interesting insight that even though the number of accidents do not increase exponentially under the presence of rain, fog and snow, these weather condition surely increase the severity of the accidents that occur under such low-visibility conditions.

Q3. Accidents caused by the percentage of humidity and air pressure rise.

As during summer season, humidity and air pressure rises causing uneasy environment and discomfort in physical activities such as driving, people loose calm and feel anger for dealing with such humid times resulting into distractions.
The correlation among these filters can be seen in data points as shown below.

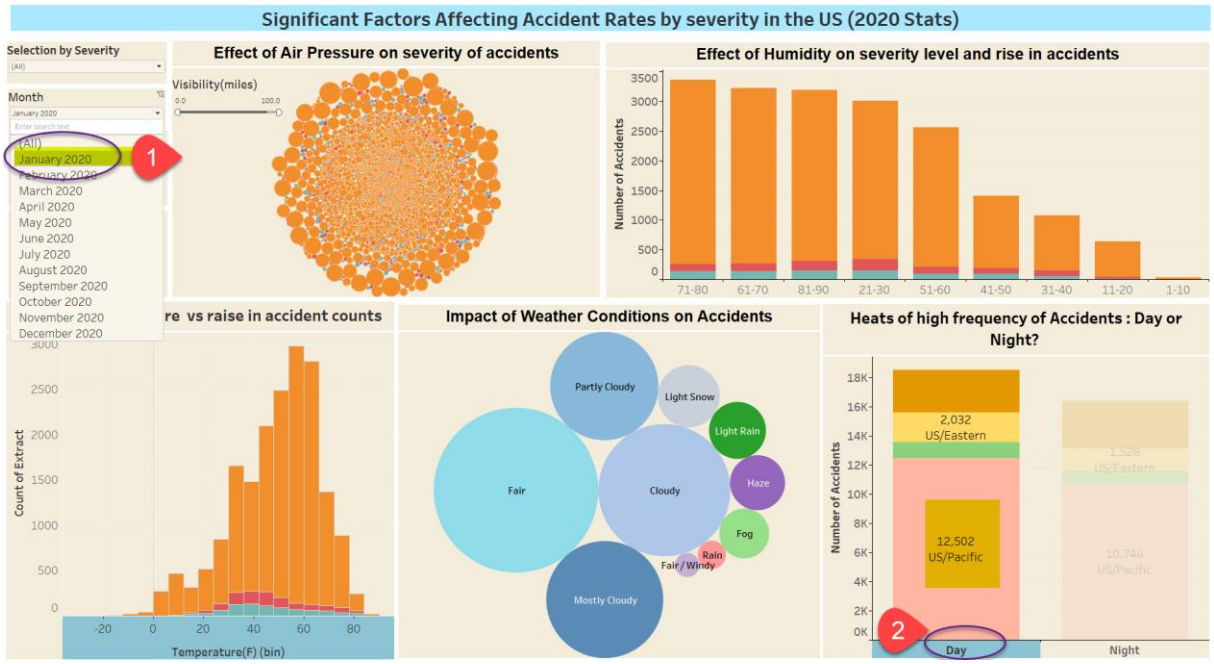
Insights drawn:

➔ We can see that In June month, the air pressure range along the number of accidents has been distributed across the severity levels of accidents and humidity peaks of 81-90 group shows although the highest number of accidents happen in 40 – 60 % of humidity, small increase in humidity causes increase possibility in accidents.



Q4. Accidents distribution over the day and night conditions.

For January 2020 month, we want to see how many accidents happened during the day.
Select drill down January from Month filter drop down list and click on “Day” axis.
The tool tip will show the details about the selection on the hover area.



Insights drawn:

- ➔ As we select the filters, we can see the data changing and telling us the story behind the accident figures. In daytime, there is more humidity as nights are comparatively much cooler. We can select specific time zones and look into more granular details of the data.

- ➔ The granularity of information will be followed in the same sequence of filters being applied in the dashboard.

- ➔ More authorized checking of automobile drivers related the alcohol consumption is required during the day as one of the factors responsible for carelessness of drivers resulting into more accidents during day and night.

Q5. Accidents distribution by severity and the temperature recorded in Fahrenheit.

Number of accidents by different severity levels namely 1,2,3,4 are being observed based on the filtered range value from the dropdown for temperature as shown below. Bins have been created to specify the values of Temperature with below settings:

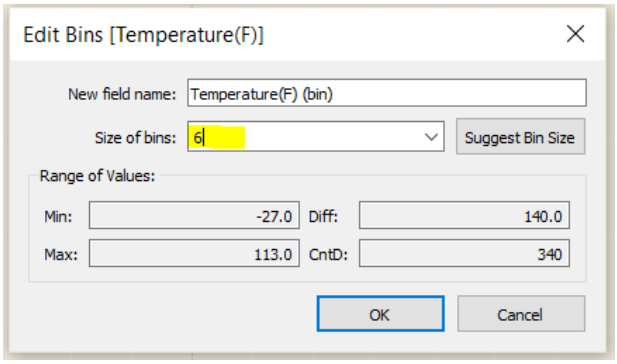


Figure- Creation of Bins on measure "Temperature(F)"

We can use the range slider for Temperature to see the changes along the dashboard and select the severity level by clicking on the histogram pallets as shown in figure below.

Case: Select 4 values for Temperature 48,54,60,66,72 to compare the distribution of accidents along with severity and choose the highest aggregated value for the bars shown – in this case "54 F" has highest count of accidents, rest of the visualizations will show the updated figures for the accidents happened at 54 F category of temperature.

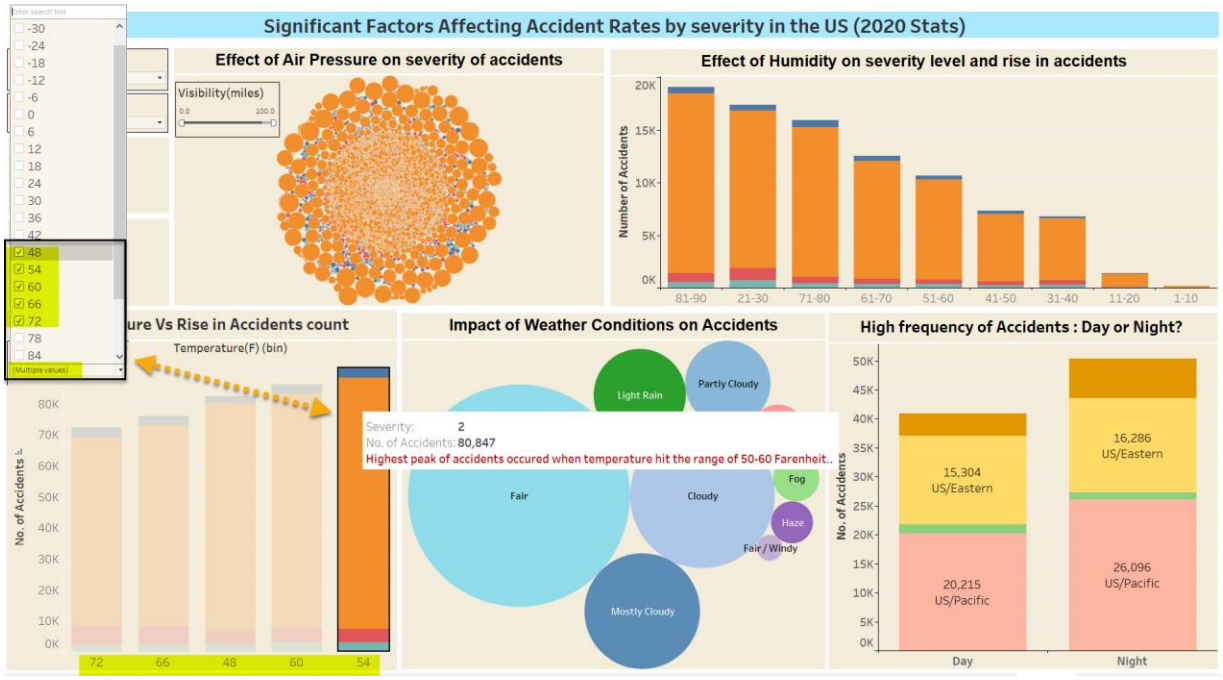
Insights drawn:

➔ Multiple questions on different levels of granularity of information can be covered here. Different environmental factors can be analyzed using different combinations such as When the temperature in 50 F – 60 F range, highest count of accidents fall in which category of humidity %.

➔ As the basic exploration of dependent factors is done, there are still many other questions which are needed to be answered to take concrete measures. That's where geographical demonstration of diagnostic analysis comes into play which can further categorize these factors based on the location.

➔ The Temperature(F) is among the less significant attributes, so the effect of temperature cannot be concluded solitary as a strong consideration for increase in accident count.

➔ Temperature can be taken as a supportive category influencing the other high relevance attributes like Humidity % and Weather condition and can be used for optimization process for comparing different combinations of variables.



1.7 Deliverable 2 – Storyboard

1.7.1 Business Scenario

Government authority's personnel needs to know where the problem is occurring and till what extent. The geographical presentation of the problem and its size is needed which includes the factors affect on states, cities, and specific accident- prone zip codes for the top cities having higher accident rate among all.

To downsize the problem area and to start from most urgent attention-needed States, the investigation for Top 10 States and their top accident-prone cities and Zip codes are needed to find out.

To explore the scope of problem statement, there are few more factors associated with transportation administrative process to segregate the stats of accident counts with respect to State level data for the following Point of Interests:

1. **Traffic Signal**
2. **Stop Sign**
3. **Crossings**
4. **Junctions**

Administrative policies and Human Resources (Shared responsibility for accident handling) Policies needs to be revised and the initial practices are required to investigate which might be misleading the roadside safety measures regarding the above stated attributes.

1.7.2 Applicable Questions

After carefully exploring all the attributes individually and with combination of significant attributes, the next step is to determine the metrics which can differentiate the problem areas at next granular level of detail as part of the diagnostic analysis process. The following questions needs to be answered to suggest recommendations at State level by the Government authorities.

Q1. Which are the most and least accident-prone States in the US.

Q2. Accident count per month of the year in the US.

Q3. Accident count per day of the week in the US.

Q4. Accident count by the hour of the day in the US.

Q5. Twenty most accident-prone Zip codes in overall US.

Q6. Representation of percentage of accidents by ageing i.e., how long the accident affected the traffic flow near the accident location. Ageing duration categories are “days”, “hours” and “minutes”.

Q7. Accident count per state using severities from low to high in the US.

Q8. Problem Analysis for the highest accident hotspot locations for most vulnerable state and city across US.

Q9. Follow up problem analysis requirement: What are the other factors contributing for increase accident rate for the most accident-prone location identified in the previous step of analysis?

Q10. What is the distribution around the Point of Interest categories shared in the data for the US accidents and which are most significant contributing into rise in accident rates?

➔ **The significance of above listed questions is relevant for the officials responsible to take corrective actions as the following things can be determined from the analysis:**

1. More concrete and immediate actions can be taken to the most accident-prone states avoiding the overall loss by minimizing the high risk of accident in those hotspot locations

2. Specific divisions can be alerted based on the holiday/vacation season when more travelling happens. For example- November- December are the months more at risk as Thanksgiving, Christmas and New-year occasions engage more people in travelling.

3. What are the safe timings to travel specifically by day and which hours are most reliable to avoid accidents can be discovered?

4. Ageing of accidents can be used for overall distribution and reporting purposes like % of accidents by minutes, hours and days based on the start_time and end_time attributes of the collected data.

5. According to the most severe accidents hotspot zip codes in most affected cities and states, special consideration to investigate the underlying cause can be investigated.

Link for Tableau Public Storyboard:

https://public.tableau.com/views/USAccidentsAnalysis2020/Storyboard?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

1.7.3 Procedure (Analysis & Visualizations)

➔ **Definition**: A Story in Tableau is a sequence of visualizations (Worksheets, or Dashboards) that work together to convey a message about what are the reasons behind the asked business problem. Each individual sheet in a story is called a story point.

Story boards are generally self-explanatory as the text and labels added via the sheets serve the purpose of finding the insights but in case the story card is based on a dashboard, a bit of explanation for navigating for the desired results is beneficial.

➔ **Storyboard elements:** US Car accident analysis storyboard consists of 12 card comprising multiple story points.
➔ **Story Cover:** Storyboard first card should consist of the objective of the analysis; in this case it is shown below:

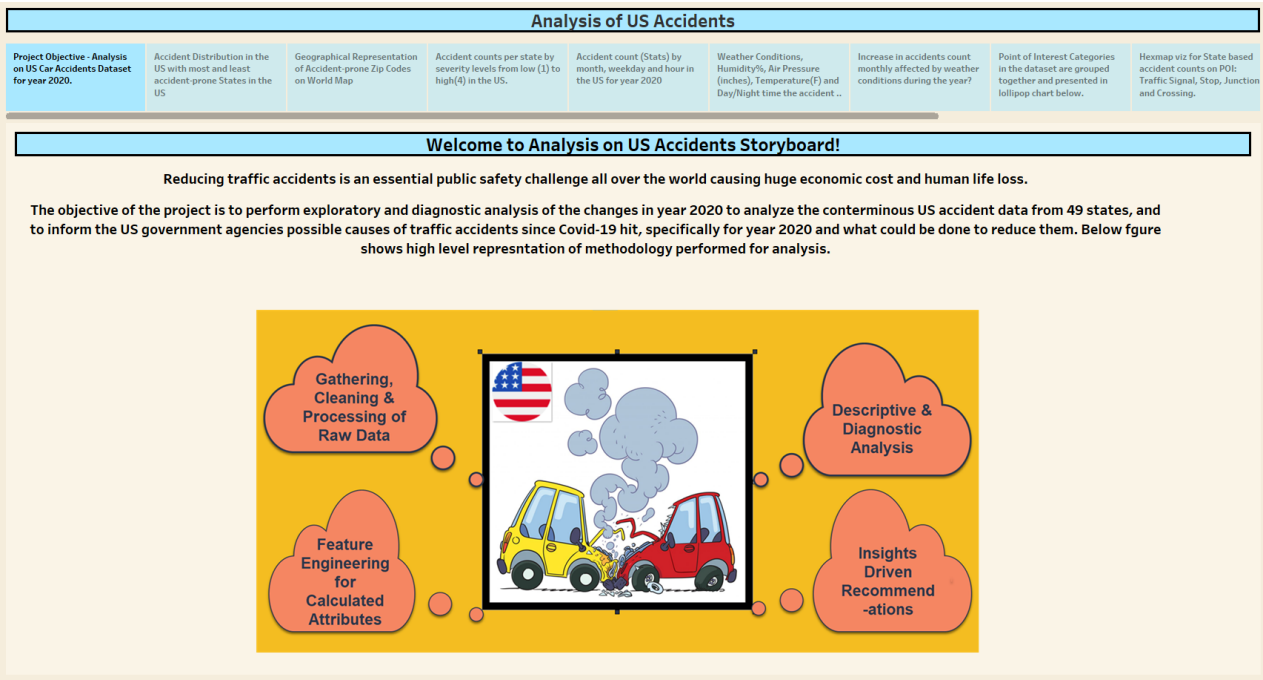


Figure 1-7 Storyboard Cover

The objective is represented through visualization which explains the methodology followed to perform the analysis. The image shows the initial data collection, data cleaning and processing of raw data to create new columns to analyze the business problems using feature engineering step. The mentioned steps are performed as an input to obtain the outcome of insights through descriptive and diagnostic analysis. Based on the insights gained from the analysis, recommendations would be provided for addressing the concern of the business issues identified.

Que. Which are the most and least accident-prone States in the US.

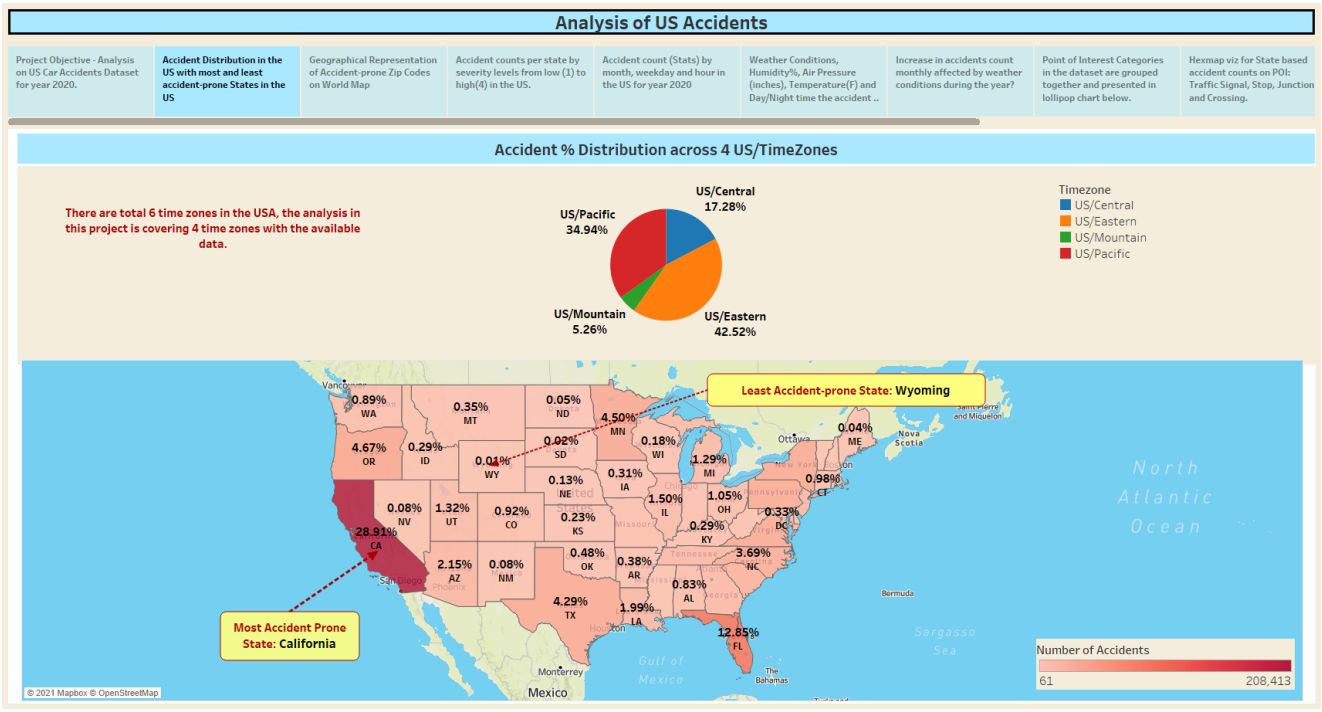


Figure 1-8 Story2

As there are 6 time zones in the US, our datapoints for this analysis falls in four time zones as presented via the pie chart visualization. This story point helps in understanding the accident count distribution in percentage across the following US/Time Zones:

1. US/Eastern – 42.52 %
2. US/Pacific – 34.94 %
3. US/Central – 17.28%
4. US/Mountain – 5.26%

The story points include the business question to show the least accident-prone State i.e., **Wyoming (0.01 %)** and most accident-prone State i.e., **California (28.91%)**. The annotation point setting is used to highlight the geographical position of two states. The color pallet used shows the dense colors for the states where more accidents have occurred in year 2020.

The background map selection used for this viz. is “Streets”. The tooltip consists of the state name along with the accident occurrence % for the highlighted state abbreviation respectively.

There are no filters used for this story as it is simply the descriptive data visualization.

Que. What is the distribution around the Point of Interest categories shared in the data for the US accidents and which are most significant contributing into rise in accident rates?

There are 13 different POI types declared in the accident dataset. Using the below calculated field, we have segregated all categories into the 1 as shown below:

POI_Type | Output Extract

```
IF  [Amenity] == TRUE THEN "Amenity"
ELSEIF [Bump] == TRUE THEN "Bump"
ELSEIF [Crossing] == TRUE THEN "Crossing"
ELSEIF [Give Way] == TRUE THEN "Give_Way"
ELSEIF [Junction] == TRUE THEN "Junction"
ELSEIF [No Exit] == TRUE THEN "No_Exit"
ELSEIF [Railway] == TRUE THEN "Railway"
ELSEIF [Roundabout] == TRUE THEN "Roundabout"
ELSEIF [Station] == TRUE THEN "Station"
ELSEIF [Stop] == TRUE THEN "Stop"
ELSEIF [Traffic Calming] == TRUE THEN "Traffic_Calming"
ELSEIF [Traffic Signal] == TRUE THEN "Traffic_Signal"
ELSEIF [Turning Loop] == TRUE THEN "Turning_Loop"
ELSE "POI_Absent"
END
```

Based on the below chart, we got the topmost POI_Types contributing to higher accident counts. There are huge number of accident records where the POI_Type value was null which is represented as “Absent” category in POI_Types.

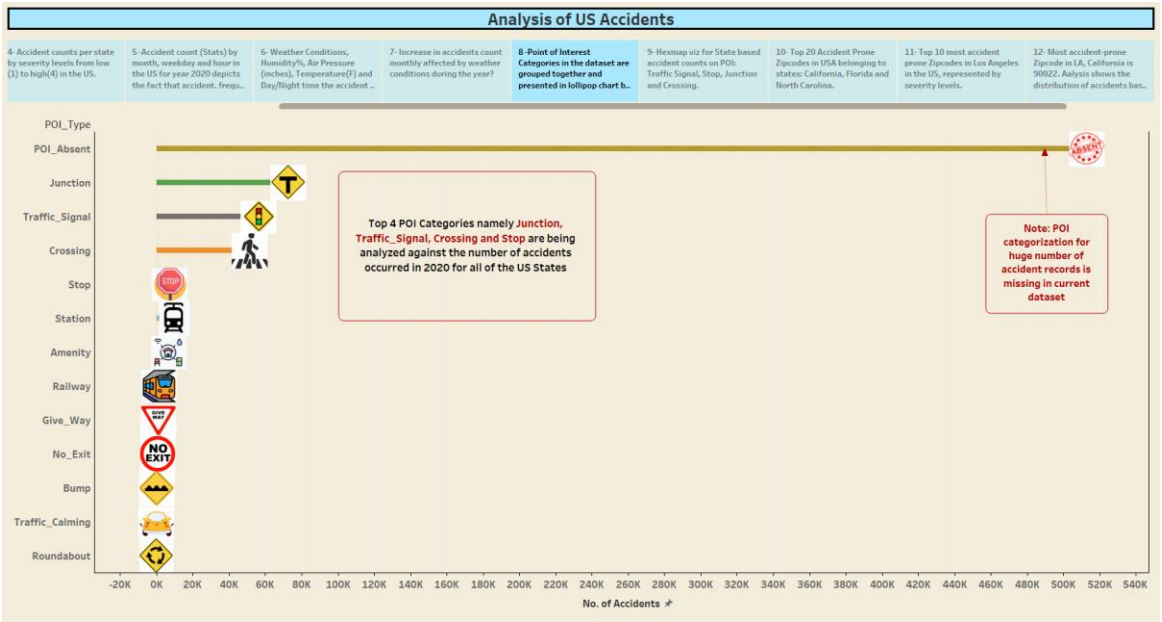


Figure- POI_Types

So, the top 4 POI_Types categories are Junction, Traffic_Signal, Stop and Crossing. These hold significant amount of weightage among all other variables resulting into higher accident rates.

Que. Accident count (Stats) by month, day of the week and hour in the US for year 2020?

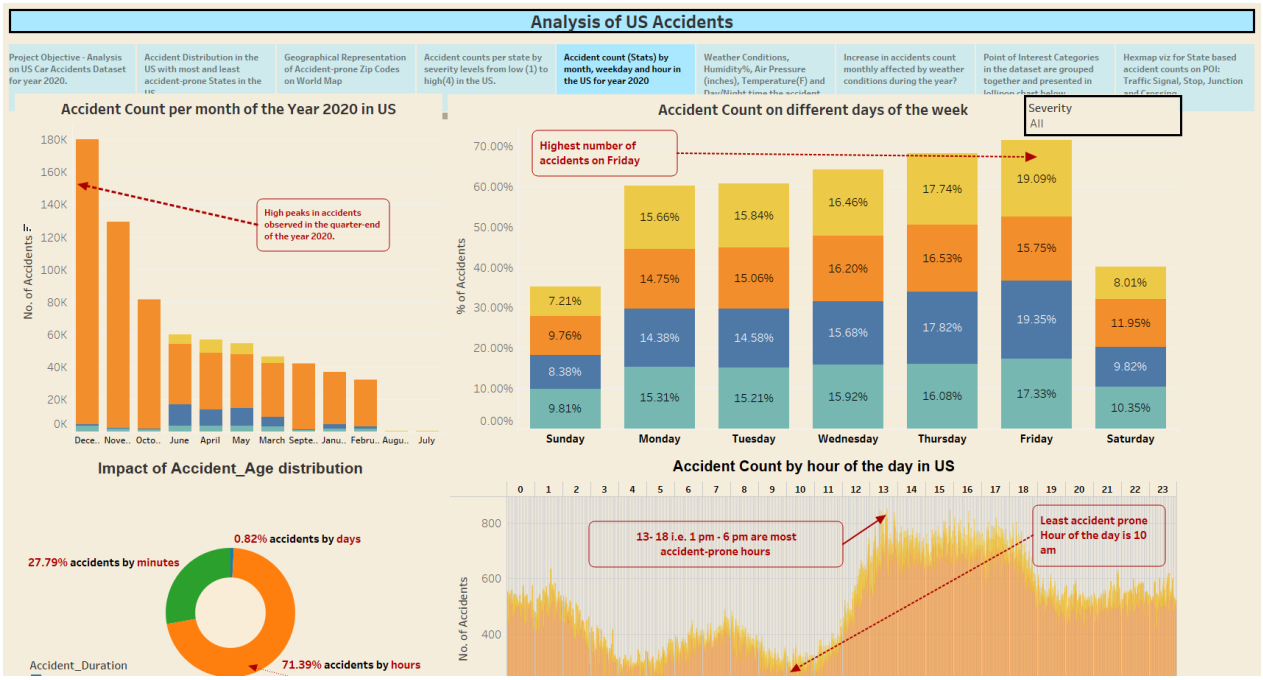


Figure – Dashboard 1

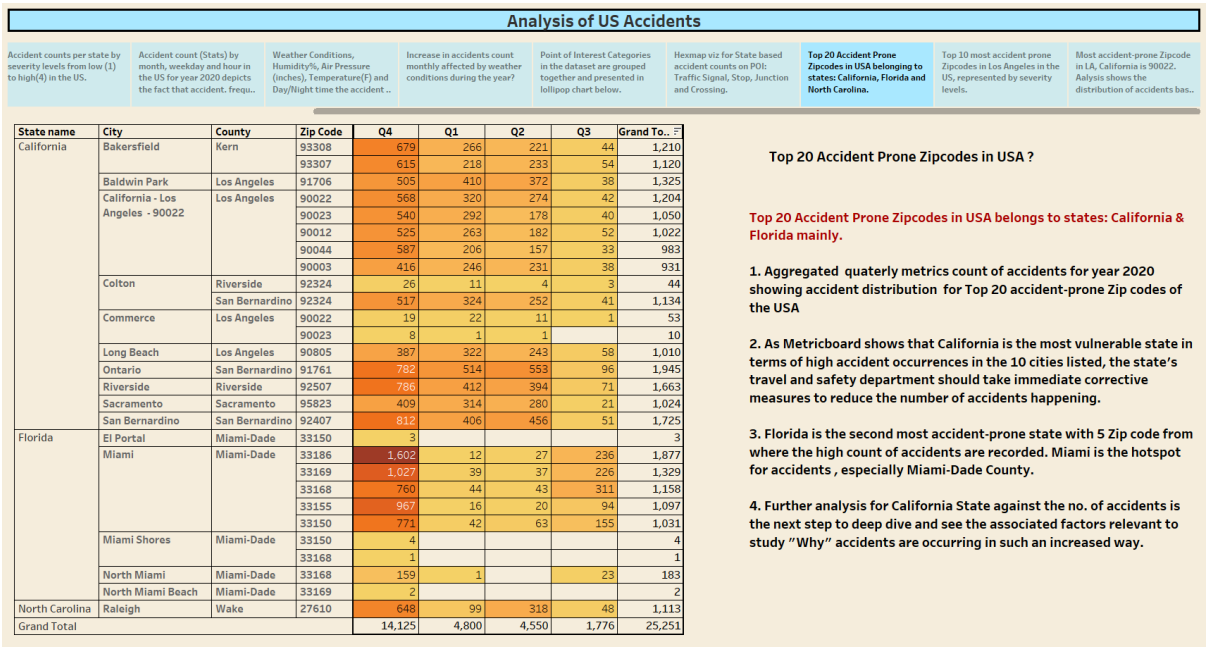
Based on the severity filter, visualizations can vary. For the default setting select as "All" i.e., Severity: 1,2,3,4 – the data is represented in the above screenshot. The filter of severity attribute is applicable to the 3 charts excluding the doughnut chart which represents the overall percentage distribution of ageing accidents categorized by minutes, hours, and days.

Insights drawn:

- ➔ Throughout the year, December has the highest count of accidents (majority with severity 2).
- ➔ Friday is the most accident-prone day of the week when approximately all severity levels of accidents occurred.
- ➔ Weekends are safer to travel as there is higher frequency of accidents over the weekdays.
- ➔ Most vulnerable hours of the day are in the range of 13.00 pm – 18.00 pm which is the later afternoon and evening time.
- ➔ Safest time to travel is considered in morning 10.00 am during the day. The time span in late night and early morning is considered safe (between 3.30 am to 5 am), assuming there are less people travelling during the night.

Que. Twenty most accident-prone Zip codes in overall US.

The following viz. shows the top 20 zip codes in the US divided into quarterly metrics of year 2020.



State name	City	County	Zip Code	Q4	Q1	Q2	Q3	Grand To..
California	Bakersfield	Kern	93308	679	266	221	44	1,210
			93307	615	218	233	54	1,120
	Baldwin Park	Los Angeles	91706	505	410	372	38	1,325
	California - Los Angeles - 90022	Los Angeles	90022	568	320	274	42	1,204
			90023	540	292	178	40	1,050
			90012	525	263	182	52	1,022
			90044	587	206	157	33	983
			90003	416	246	231	38	931
	Colton	Riverside	92324	26	11	4	3	44
		San Bernardino	92324	517	324	252	41	1,134
	Commerce	Los Angeles	90022	19	22	11	1	53
			90023	8	1	1		10
	Long Beach	Los Angeles	90805	387	322	243	58	1,010
	Ontario	San Bernardino	91761	782	514	553	96	1,945
	Riverside	Riverside	92507	786	412	394	71	1,663
	Sacramento	Sacramento	95823	409	314	280	21	1,024
	San Bernardino	San Bernardino	92407	812	406	456	51	1,725
Florida	El Portal	Miami-Dade	33150	3				3
	Miami	Miami-Dade	33186	1,602	12	27	236	1,877
			33169	1,027	39	37	226	1,329
			33168	760	44	43	311	1,158
			33155	967	16	20	94	1,097
			33150	771	42	63	155	1,031
	Miami Shores	Miami-Dade	33150	4				4
			33168	1				1
	North Miami	Miami-Dade	33168	159	1		23	183
	North Miami Beach	Miami-Dade	33169	2				2
North Carolina	Raleigh	Wake	27610	648	99	318	48	1,113
Grand Total				14,125	4,800	4,550	1,776	25,251

Figure- Tabular representation viz

The top accident-prone zip codes are listed, and their respective City, County and States information is also included in the table shown.

Insights drawn:

- ➔ California and Florida are the most accident-prone states who needs immediate attention to address the problem of high accident rate across the country.
- ➔ California is containing 10 zip codes having high accident counts where huge accidents have occurred in Los Angeles City followed by Miami, Florida with Miami-Dade as the hotspot for high accidents occurrences.
- ➔ The grand totals shows that these accidents needs to be analyzed further with respect to other combination of variables responsible for accidents.

Que. Accidents count per state using severities from low to high in the US.

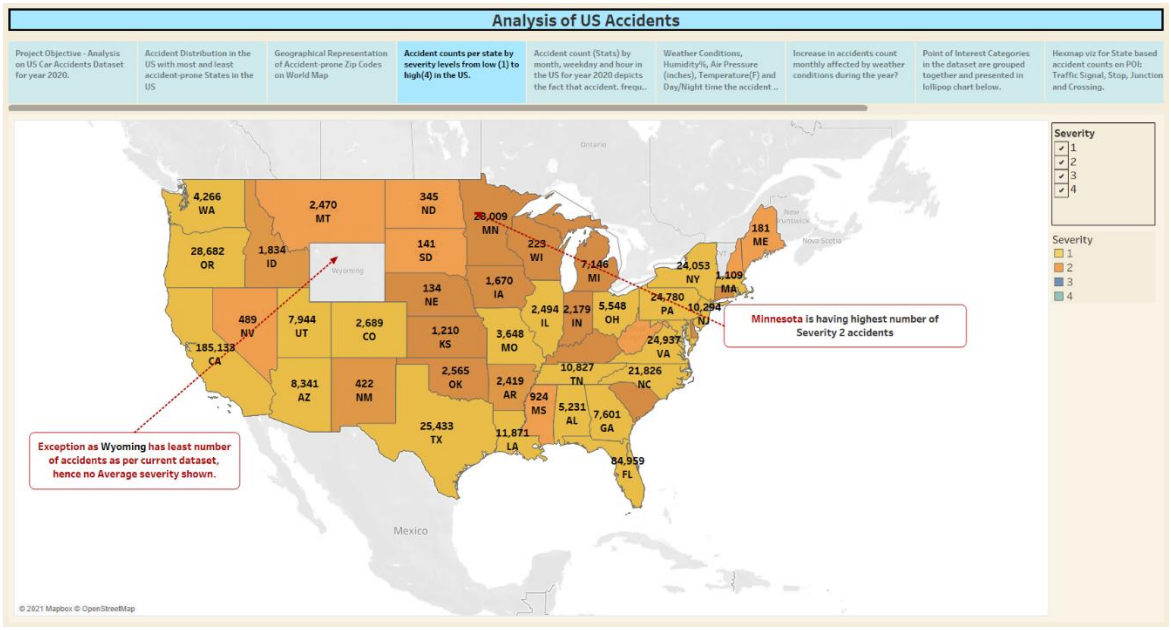


Figure- Map by Accident Severity

➔ Using the filter for severity, different visualizations can be achieved to locate the high severity states. As shown in the above map screenshot, Wyoming is not having any color differentiation as the number of accidents are too low compared to other states metrics.

➔ Minnesota is having highest number of severity 2 incidents i.e., 28,009 to be precise.

➔ Although Florida is having 84,959 accidents count but the severity attribute tells us about how much delay will be imposed on the traffic due to accident occurrence in nearby routes.

➔ By finding the highest severity accidents prone state, federal administrative officials can take a look into the problem and start planning on how to correct the mis-happenings in the system/policies/administration handling infrastructure.

➔ To find out in case suppose, severity 4 accidents distribution for the states, Click on the severity filter as shown in below figure.

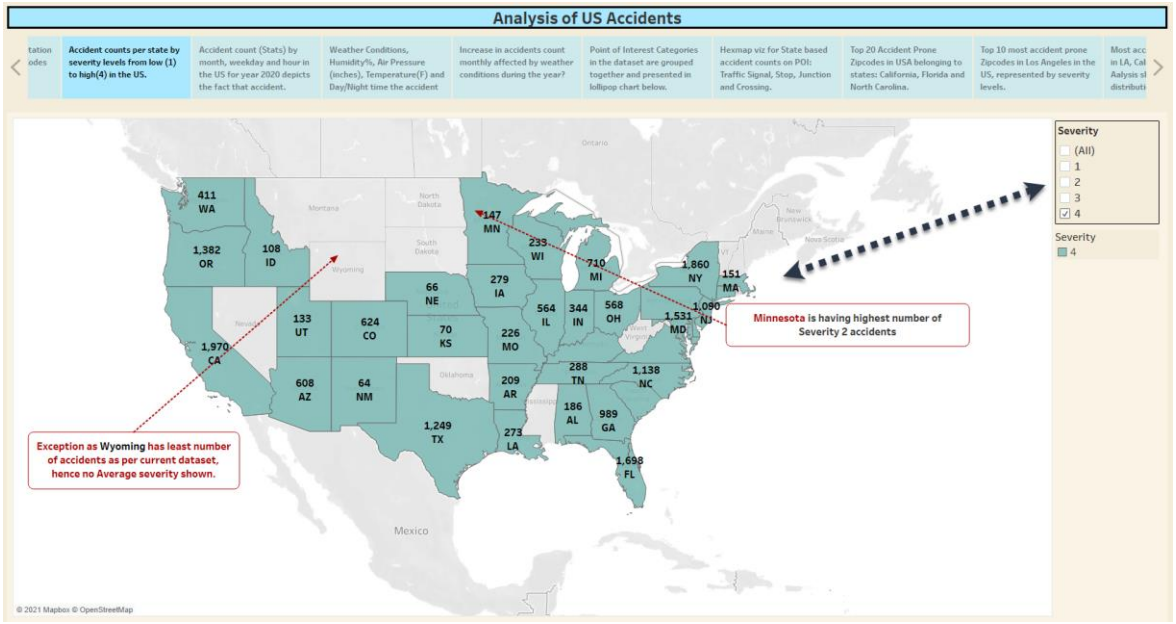


Figure – Map (Filter Demo)

Que. Problem Analysis for the highest accident hotspot locations for most vulnerable state and city across US with severity distribution.

➔ Topmost accident-prone hotspot is 90022 as number of accidents are huge (1,259) for this location. It is noteworthy that there are no severity-1 accidents and majority of the accidents are of Severity-2 causing short delay over the traffic in all specified locations.

➔ The story represents the next granular level to reach the root cause of the problem for high accident counts in Los Angeles in these specific zip code areas as part of the diagnostic analysis in the next story.

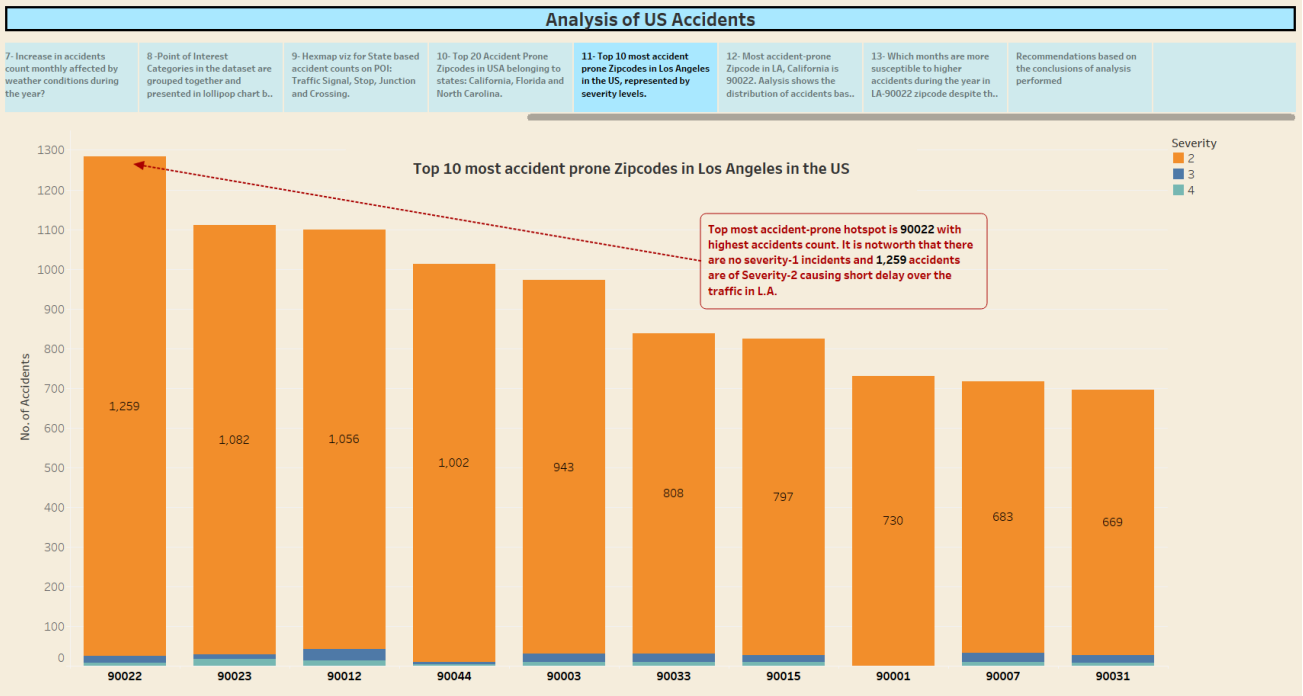


Figure- Top 10 Zip codes in L.A. California, US

Que. Follow up problem analysis requirement: What are the other factors contributing for increase accident rate for the most accident-prone location identified in the previous step of analysis?

Factors to be analyzed for jotting down the root cause pain points:

1. Time of accident occurrence i.e., day/night with respect to weather conditions.
2. As most accidents happen during fair weather conditions, how good weather criteria can be concluded to know accident counts per month for LA-90022 California for year 2020 ?

The below figures represents the answers to the above listed questions in order to conclude our analysis.

Insights Drawn:

➔ Through the bar chart for top 10 highly contributing weather conditions in US against the number of accidents in zip code 90022, it is proved that the distribution of accidents is more when the weather is fair pointing that is the most preferred travelling time as in general.

➔ More accidents involve night-time travelling.

➔ The interesting question is that California is known for the Clear/fair weather for most of the months throughout the year. Still, accidents are occurring. The peak count of accidents per month in clear weather is investigated in next part of the story point.

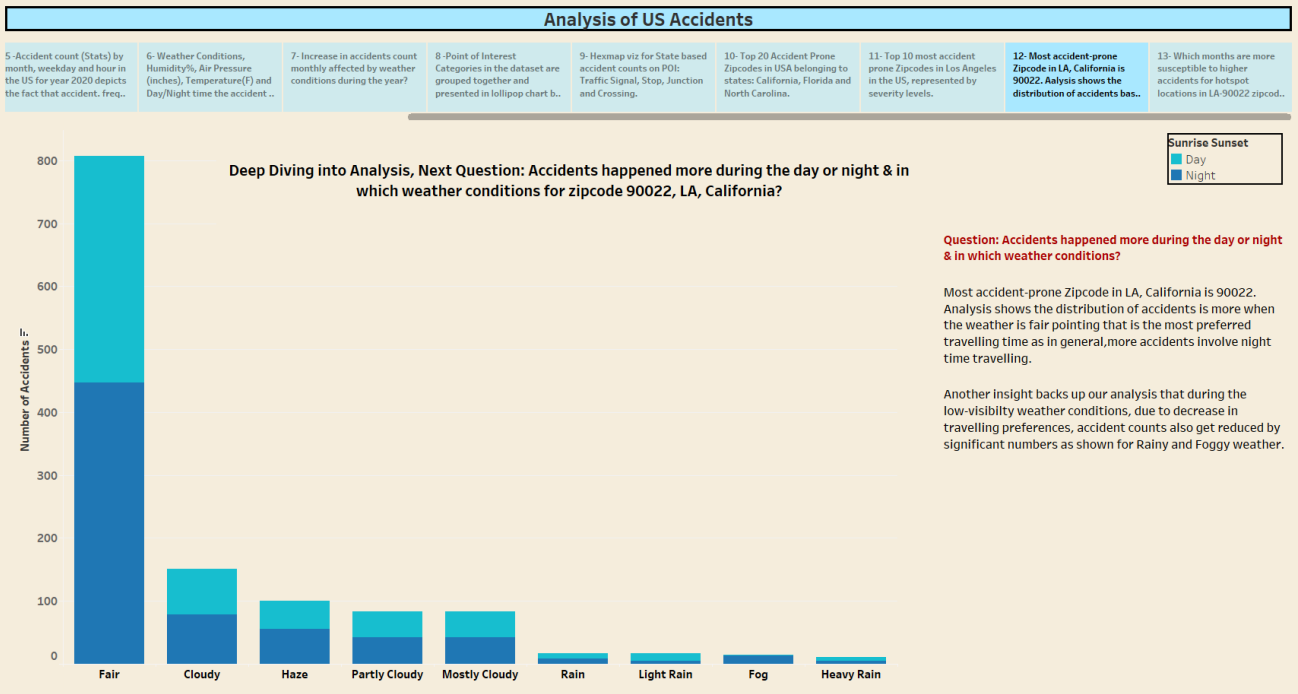


Figure – Why 90022 Zip code is hotspot for accidents?

Which months are more susceptible to higher accidents during the year in LA-90022 zip code despite the clear weather?

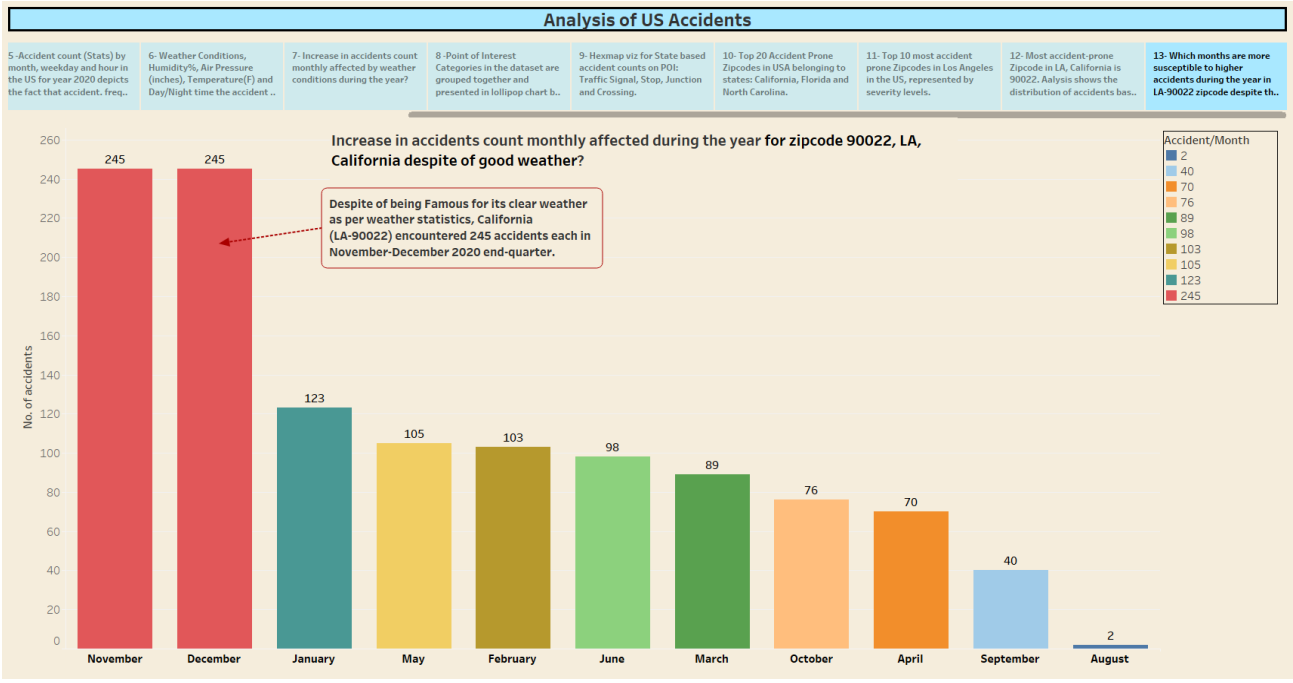


Figure – 90022- LA accidents per month count

Using feature engineering , created a new calculated measure to get the **aggregated count monthly for accidents occurring in fair weather conditions**. Refer the below screenshot for the details.

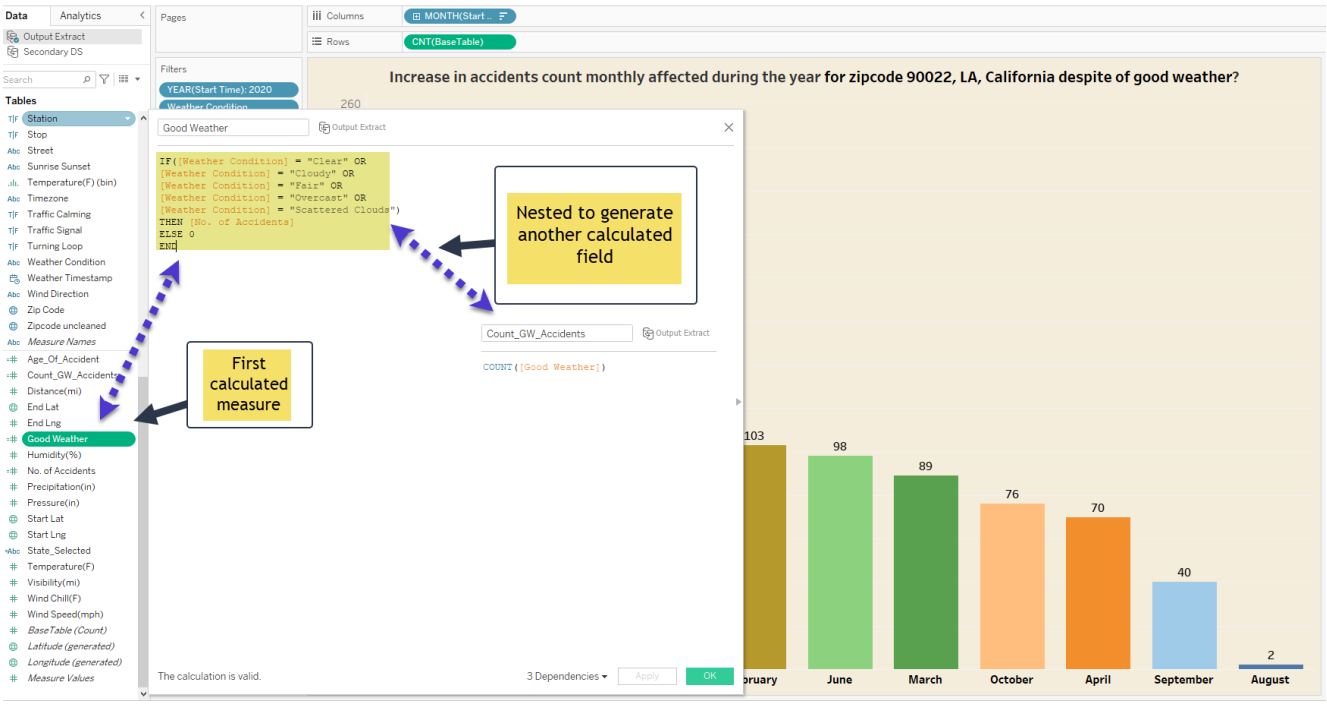


Figure – Feature Engineering: "Good Weather"

POI_Type categories by "State" selection "California" metrics are shown in below dashboard:

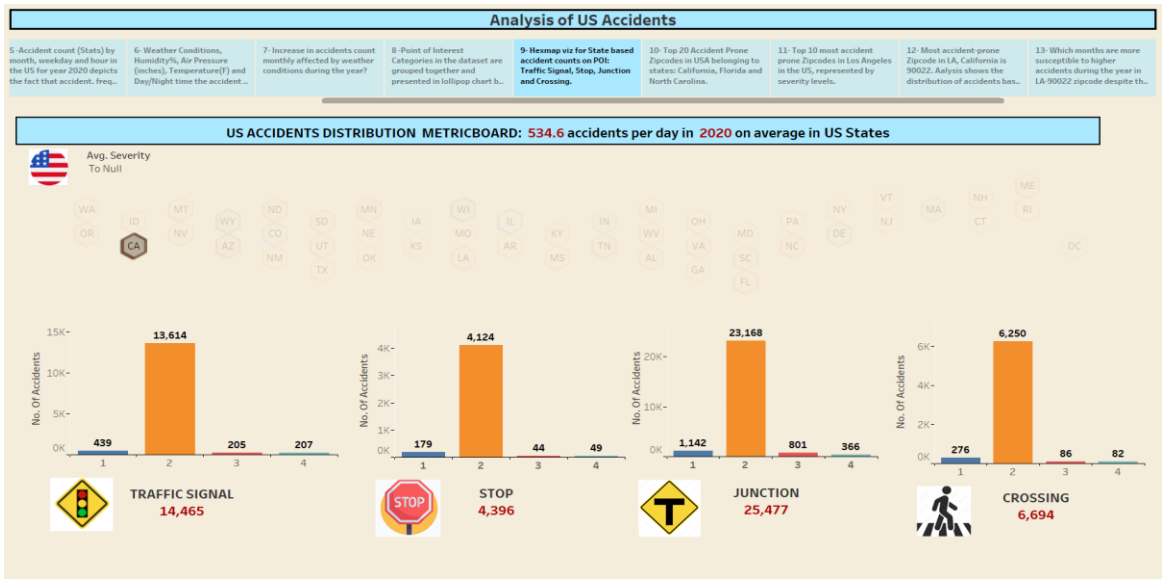


Figure- Hex Map Dashboard filtered for CA state POI_Type metric distribution

1.8 Conclusions

→ The significant findings from the overall accident dataset analysis are:

(a) End quarter months (October, November, December) are having huge accidents count per month as compared to initial quarters.

(b) Nighttime is safer to travel, 3 am to 6 am to be precise.

(c) Most of the accidents are happening in fair weather conditions, especially in accident-prone states.

(d) Weather, humidity(%), Pressure (air pressure in inches), and location are the factors responsible for 35% of the accidents.

(e) Weekdays are more susceptible to encounter more accidents as compared to weekends.

(f) Severity of accidents shows that even Severity-2 i.e., short delay causing accidents also result in huge traffic/congestion issues blocking other routes making them more likely to have accidents.

→ Findings specific to California State accident prone city and zip code:

(a) Top 10 accident-prone cities across US belong to California state with 10 hotspots zip codes, out of which Los Angeles holds big numbers.

(b) Florida is the second state vulnerable to high accident rates. Miami is the main culprit for such accidents metrics as “Miami-Dade” county is where 10 zip codes have come into picture.

(c) Top POI_Type categories are factors responsible for 15% of the accidents in California.

1.8.1 Recommendations

The main recommendations from the project focus are on Policy, Administrative, and Human behavior-related changes that can be implemented by the state and the federal government. As there are other relevant factors apart from used this in dataset, it is better to gather the information from NHTSA for speed, Driver’s data, Age of the driver and victims in the accident involved, road condition, etc.

Based on the current analysis and the insights deduced, following seems to be necessary at this point in time:

1. As we can see, most of the accidents occur in clear and fair weather which indicate that most of the accidents might be the result of negligence of the driver. Cloudy skies can make driving more dangerous because it decreases visibility. Also, cloudy skies make it harder to see potholes, black eyes, and plowed roads. Precautionary alerts should be sent to warn drivers about cloudy/ overcast conditions well in advance as significant number of accidents are reported in such

weather conditions. Deep-dive analysis is needed to find the exact reasons and backing up our thesis/assumptions.

2. Over-crowding and road congestions should be avoided which happen due to more accidents belonging to severity levels.

3. Proper traffic/congestion alert signals, side boards, informatory signs, LED indicators can be installed to alert drivers while approaching the POI types mentioned such as Stop signs, Traffic Signals, Crossings, Junctions, and other possible crowded/accident-prone places.

4. Traffic safety rules and best practices guidelines should be included in educating people opting for driving licenses, especially in highly affected hotspots to implement safe practices.

5. Strick rules and regulations should be followed and implemented by the administration about following speed limits in identified accident prone cities.

6. Precautionary alerts should be sent to warn people about weather conditions getting worse for travelling and reconsideration for planning ahead as significant no. of accidents happen in low visibility conditions.

7. Strick measures should be taken to decrease the possible conditions (which are under human-behavior control) , identified during the day especially on the weekdays.

1.9 Summary Table (Learning Objectives)

Learning Objectives	Activities/Tasks	How would you measure the reader's learning
Data Exploratory Analysis	Carefully performed Data Preparation Procedure using Tableau Prep and cleaned the dataset to eliminate anomalies.	With step by step click through screenshots and elaboration of tasks to be performed were well documented for better understanding of reader's learning and to justify the actions taken for the analysis.
	All variables were explored based on the basic meaning and the data points for the US Accident dataset with respect to the business scenarios mentioned.	Individual attributes were considered to demonstrate the correlation among the attributes used. Significant analysis by running the regression models using R/Python was performed. To obtain the attributes with most relevance to the problem statement, the aggregated list was presented in the analysis & visualizations section of the document.
	Documentation carried out for all the actions taken to explore the US Accidents dataset with handling anomalies. Run the exploration, view the changed stats for the problem questions being analyzed based on the selected indicator attributes.	Dashboard representing all the possible granularities of the data. Reader can navigate through the filter selections for different visualizations and understand the insights based on slicing and dicing performed on the data. All the instructions to operate the dashboard are clearly explained in the documentation as well.
Descriptive Analysis	Combination of attributes were analyzed to deduce the insights via data understanding and feature engineering for deep diving to find the root cause of the issues	Goal achieved and a list of significant variables was taken to perform the diagnostic analysis. Updated Statistics will be visualized ready to be used for further steps for the reader through the storyboard consisting of the dashboard viz.
Diagnostic Analysis	Identification of KPI's and making of visualizations, applications evaluated for the list of identified KPIs	By following the tasks documented, readers can analyze and obtain the KPIs and understand how the factors represented are affecting the increase in accident count.
	Run exploration, chose a particular Severity/Weather_Condition/Month/Hour/Day of the year 2020 for the given dataset and view the results in the dashboard.	Reader can understand the different pain points and problem areas of accidents occurrences. The best time to travel, which Weather_Condition most affects the raised counts of accidents, Top 20 Zip codes across the country with highest accident rates, Which City, State and Zip codes are hotspot for immediate consideration for government officials, etc.
	Insights shared based on the analysis findings	Reader can use the insights and recommendations proposed as solutions to the problem areas and understand the necessary corrective measures to be taken to resolve/minimize the accidents.

1.10 Limitations

Due to information constraint, part of the original dataset was removed which consisted of 3million records in the dataset for US Car Accidents (2019). After removing the sensitive information and attributes, updated dataset "US Accidents (Updated 2020)" was created and shared on Kaggle for academic and research purposes.

**Disclaimer: It is a valid consideration to keep in mind that it is possible to find false trends because the data could actually be missing information on some crashes that did not end up being reported or recorded.*

1.10.1 Recommended Citations/Acknowledgements

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

1.10.2 Usage Policy and Legal Disclaimer

This dataset is being distributed only for Research purposes, under Creative Commons Attribution-Noncommercial-Share Alike license (CC BY-NC-SA 4.0). By downloading this dataset, user agrees to use this data only for non-commercial, research, or academic applications.

1.11 References

1. Over 100 Car Accident Statistics for 2020 | U.S. and Global (safer-america.com)
2. Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights (arxiv.org)
3. Road Safety Facts — Association for Safe International Road Travel (asirt.org)
4. NHTSA | National Highway Traffic Safety Administration
5. Causes and Solutions of Car Accidents - 1117 Words | Report Example (ivypanda.com)
6. Contiguous United States - Wikipedia
7. https://scholarworks.lib.csusb.edu/etd/979
8. [1906.05409] A Countrywide Traffic Accident Dataset (arxiv.org)
9. US-Accidents Dataset | Papers With Code
10. HiRes2-1084x1035.jpg (1084×1035) (sagoomotors.co.ke)
11. 7340cbb107d570f0260bc23746c47925.jpg (5198×4022) (pinimg.com)